UNIFORM POST SELECTION INFERENCE FOR LAD REGRESSION MODELS

A. BELLONI, V. CHERNOZHUKOV, AND K. KATO

Abstract. We develop uniformly valid confidence regions for a regression coefficient in a high-dimensional sparse LAD (least absolute deviation or median) regression model. The setting is one where the number of regressors p could be large in comparison to the sample size n, but only $s \ll n$ of them are needed to accurately describe the regression function. Our new methods are based on the instrumental LAD regression estimator that assembles the optimal estimating equation from either post ℓ_1-penalized LAD regression or ℓ_1-penalized LAD regression. The estimating equation is immunized against non-regular estimation of nuisance part of the regression function, in the sense of Neyman. We establish that in a homoscedastic regression model, under certain conditions, the instrumental LAD regression estimator of the regression coefficient is asymptotically root-n normal uniformly with respect to the underlying sparse model. The resulting confidence regions are valid uniformly with respect to the underlying model. The new inference methods outperform the naive, "oracle based" inference methods, which are known to be not uniformly valid – with coverage property failing to hold uniformly with respect the underlying model – even in the setting with $p = 2$. We also provide Monte-Carlo experiments which demonstrate that standard post-selection inference breaks down over large parts of the parameter space, and the proposed method does not.

Key words: median regression, uniformly valid inference, instruments, Neymanization, optimality, sparsity, post selection inference

1. Introduction

We consider the following regression model

$$y_i = d_i\alpha_0 + x_i'\beta_0 + \epsilon_i, \; i = 1, \ldots, n, \tag{1.1}$$

where d_i is the "main regressor" of interest, whose coefficient α_0 we would like to estimate and perform (robust) inference on. The $(x_i)_{i=1}^n$ are other high-dimensional regressors or "controls" and are treated as fixed (d_i's are random). The regression error ϵ_i is independent of d_i and has median 0. The errors $(\epsilon_i)_{i=1}^n$ are i.i.d. with distribution function $F(\cdot)$ and probability density function $f_\epsilon(\cdot)$ such that $F(0) = 1/2$ and $f_\epsilon = f_\epsilon(0) > 0$. The assumption on the error term motivates the use of the least absolute deviation (LAD) or median regression, suitably adjusted for use in high-dimensional settings.

Date: First version: May 2012, this version April 30, 2013. We would like to thank the participants of Luminy conference on Nonparametric and high-dimensional statistics (December 2012), Oberwolfach workshop on Frontiers in Quantile Regression (November 2012), 8th World Congress in Probability and Statistics (August 2012), and seminar at the University of Michigan (October 2012). We are grateful to Sara van de Geer, Xuming He, Richard Nickl, Roger Koenker, Vladimir Koltchinskii, Steve Portnoy, Philippe Rigollet, and Bin Yu for useful comments and discussions.

The dimension p of "controls" x_i is large, potentially much larger than n, which creates a challenge for inference on α_0. Although the unknown true parameter β_0 lies in this large space, the key assumption that will make estimation possible is its sparsity, namely $T = \text{support}(\beta_0)$ has $s < n$ elements (where s can depend on n; we shall use array asymptotics). This in turn motivates the use of regularization or model selection methods.

A standard (non-robust) approach towards inference in this setting would be first to perform model selection via the ℓ_1-penalized LAD regression estimator

$$(\widehat{\alpha}, \widehat{\beta}) \in \arg\min_{\alpha,\beta} \mathbb{E}_n[|y_i - d_i\alpha - x_i'\beta|] + \frac{\lambda}{n}\|(\alpha, \beta')'\|_1, \tag{1.2}$$

and then to use the post-model selection estimator

$$(\widetilde{\alpha}, \widetilde{\beta}) \in \arg\min_{\alpha,\beta} \left\{ \mathbb{E}_n[|y_i - d_i\alpha - x_i'\beta|] : \beta_j = 0 \text{ if } \widehat{\beta}_j = 0 \right\} \tag{1.3}$$

to perform "usual" inference for α_0. (The notation $\mathbb{E}_n[\cdot]$ denotes the average over index $1 \leqslant i \leqslant n$.)

This standard approach is justified if (1.2) achieves perfect model selection with probability approaching 1, so that the estimator (1.3) has the "oracle" property with probability approaching 1. However conditions for "perfect selection" are very restrictive in this model, in particular, requiring significant separation of non-zero coefficients away from zero. If these conditions do not hold, the estimator $\tilde{\alpha}$ does not converge to α_0 at the $\sqrt{n}$-rate – uniformly with respect to the underlying model– which implies that "usual" inference breaks down and is not valid. (The statements continue to apply if α is not penalized in (1.2), α is restricted in (1.3), or if thresholding is applied.) We shall demonstrate the breakdown of such naive inference in the Monte-Carlo experiments where non-zero coefficients in θ_0 are not significantly separated from zero.

Note that the breakdown of inference does not mean that the aforementioned procedures are not suitable for prediction purposes. Indeed, the ℓ_1-LAD estimator (1.2) and post ℓ_1-LAD estimator (1.3) attain (essentially) optimal rates $\sqrt{(s \log p)/n}$ of convergence for estimating the entire median regression function, as has been shown in [24, 3, 13, 26] and in [3]. This property means that while these procedures will not deliver perfect model recovery, they will only make "moderate" model selection mistakes (omitting only controls with coefficients local to zero).

To achieve uniformly valid inferential performance we propose a procedure whose performance does not require perfect model selection and allows potential "moderate" model selection mistakes. The latter feature is critical in achieving uniformity over a large class of data generating processes, similarly to the results for instrumental regression and mean regression studied in [27], [2], [7], [6]. This allows us to overcome the impact of (moderate) model selection mistakes on inference, avoiding (in part) the criticisms in [17], who prove that the "oracle property" sometime achieved by the naive estimators necessarily implies the failure of uniform validity of inference and their semiparametric inefficiency [18].

In order to achieve robustness with respect to moderate model selection mistakes, it will be necessary to achieve the proper orthogonality condition between the main regressors and the control variables.

Towards that goal the following auxiliary equation plays a key role (in the homoscedastic case):

$$d_i = x_i'\theta_0 + v_i, \ \mathrm{E}[v_i] = 0, \ i = 1, \ldots, n; \tag{1.4}$$

describing the relevant dependence of the regressor of interest d_i to the other controls x_i. We shall assume the sparsity of θ_0, namely $T_d = \text{support}(\theta_0)$ has at most $s < n$ elements, and estimate the relation (1.4) via Lasso or post-Lasso methods described below.

Given v_i, which "partials out" the effect of x_i from d_i, we shall use it as an instrument in the following estimating equations for α_0:

$$\mathrm{E}[\varphi(y_i - d_i\alpha_0 - x_i'\beta_0)v_i] = 0, \quad i = 1, ..., n,$$

where $\varphi(t) = 1/2 - 1(t < 1/2)$. We shall use the empirical analog of this equation to form an instrumental LAD regression estimator of α_0, using a plug-in estimator for $x_i'\beta_0$. The estimating equation above has the following feature:

$$\frac{\partial}{\partial\beta}\mathrm{E}[\varphi(y_i - d_i\alpha_0 - x_i'\beta)v_i]\bigg|_{\beta=\beta_0} = 0, \quad i = 1, ..., n, \tag{1.5}$$

As a result, the estimator of α_0 will be "immunized" against "crude" estimation of $x_i'\beta_0$, for example, via a post-selection procedure or some regularization procedure. As we explain in Section 5, such immunization ideas can be traced back to Neyman ([19, 20]).

Our estimation procedure has the following three steps.

Step 1: Estimation of the confounding function $x_i'\beta_0$ in (1.1).
Step 2: Estimation of the instruments (residuals) v_i in (1.4).
Step 3: Estimation of the main effect α_0 based on the instrumental LAD regression using v_i as instruments for d_i.

Each step is computationally tractable, involving solutions of convex problems and a one-dimensional search, and relies on a different identification condition which in turn requires a different estimation procedure:

Step 1 constructs an estimate for the nuisance function $x_i'\beta_0$ and not an estimate for α_0. Here we do not need a $\sqrt{n}$-rate consistency for the estimates of the nuisance function; slower rate like $o(n^{-1/4})$ will suffice. Thus, this can be based either on the ℓ_1-LAD regression estimator (1.2) or the associated post-model selection estimator (1.3).

Step 2 partials out the impact of the covariates x_i on the main regressor d_i, obtaining the estimate of the residuals v_i in the decomposition (1.4). In order to estimate these residuals we rely either on heteroscedastic Lasso [2], a version of the Lasso estimator of [23, 9]:

$$\widehat{\theta} \in \arg\min_{\theta} \mathbb{E}_n[(d_i - x_i'\theta)^2] + \frac{\lambda}{n}\|\widehat{\Gamma}\theta\|_1 \text{ and set } \widehat{v}_i = d_i - x_i'\widehat{\theta}, \ i = 1, \ldots, n, \tag{1.6}$$

where λ and $\widehat{\Gamma}$ are the penalty level and data-driven penalty loadings described in [2] (restated in Appendix D), or the associated post-model selection estimator (Post-Lasso) [4, 2] defined as

$$\widetilde{\theta} \in \arg\min_{\theta} \left\{\mathbb{E}_n[(d_i - x_i'\theta)^2] : \theta_j = 0 \text{ if } \widehat{\theta}_j = 0\right\} \text{ and set } \widehat{v}_i = d_i - x_i'\widetilde{\theta}. \tag{1.7}$$

Step 3 constructs an estimator $\check{\alpha}$ of the coefficient α_0 via an instrumental LAD regression proposed in [10], using $(\widehat{v}_i)_{i=1}^n$ as instruments. Formally, $\check{\alpha}$ is defined as

$$\check{\alpha} \in \arg\inf_{\alpha \in \mathcal{A}} L_n(\alpha), \text{ where } L_n(\alpha) = \frac{4|\mathbb{E}_n[\varphi(y_i - x_i'\widehat{\beta} - d_i\alpha)\widehat{v}_i]|^2}{\mathbb{E}_n[\widehat{v}_i^2]}, \tag{1.8}$$

$\varphi(t) = 1/2 - 1\{t \leqslant 0\}$ and $\mathcal{A}$ is a parameter space for α_0. We will analyze the choice of $\mathcal{A} = [\widehat{\alpha} - C\log^{-1} n, \widehat{\alpha} + C\log^{-1} n]$ with a suitable constant $C > 0$. [1] Several other choices for $\mathcal{A}$ are possible.

Our main result establishes conditions under which $\check{\alpha}$ is root-n consistent for α_0, asymptotically normal, and achieves the semi-parametric efficiency bound for estimating α_0 in the current homoscedastic setting, provided that $(s^3 \log^3 p)/n \to 0$ and other regularity conditions hold. Specifically, we show that, despite possible model selection mistakes in Steps 1 and 2, the estimator $\check{\alpha}$ obeys

$$\sigma_n^{-1}\sqrt{n}(\check{\alpha} - \alpha_0) \rightsquigarrow N(0,1), \tag{1.9}$$

where $\sigma_n^2 := 1/(4f_\epsilon^2\bar{\mathrm{E}}[v_i^2])$ with $f_\epsilon = f_\epsilon(0)$. An alternative (and more robust) expression for σ_n^2 is given by Huber's sandwich:

$$\sigma_n^2 = J^{-1}\Omega J^{-1}, \text{ where } \Omega := \bar{\mathrm{E}}[v_i^2]/4 \text{ and } J := \bar{\mathrm{E}}[f_\epsilon d_i v_i]. \tag{1.10}$$

We recommend to estimate Ω by the plug-in method and to estimate J by Powell's method [21]. Furthermore, we show that the criterion function at the true value α_0 in Step 3 has the following pivotal behavior

$$nL_n(\alpha_0) \rightsquigarrow \chi^2(1). \tag{1.11}$$

This allows the construction of a confidence region $\widehat{A}_{n,\xi}$ with asymptotic coverage $1 - \xi$ based on the statistic L_n,

$$\mathrm{P}(\alpha_0 \in \widehat{A}_{n,\xi}) \to 1 - \xi \text{ where } \widehat{A}_{n,\xi} = \{\alpha \in \mathcal{A} : nL_n(\alpha) \leqslant (1-\xi)\text{-quantile of } \chi^2(1)\}. \tag{1.12}$$

Importantly, the robustness with respect to moderate model selection mistakes, which occurs because of (1.5), allows the results (1.9) and (1.11) to hold uniformly over a large range of data generating processes, similarly to the results for instrumental regression and partially linear mean regression model established in [6, 27, 2]. One of our proposed algorithms explicitly uses ℓ_1-regularization methods, similarly to [27] and [2], while the main algorithm we propose uses post-selection methods, similarly to [6, 2].

Throughout the paper, we use array asymptotics – asymptotics where the model changes with n – to better capture some finite-sample phenomena such as "small coefficients" that are local to zero. This ensures the robustness of conclusions with respect to perturbations of the data-generating process along various model sequences. This robustness, in turn, translates into uniform validity of confidence regions over substantial regions of data-generating processes.

[1] For numerical experiments we used $C = 10(\mathbb{E}_n[d_i^2])^{-1/2}$ and typically we normalize $\mathbb{E}_n[d_i^2] = 1$).

1.1. **Notation and convention.** Denote by $(\Omega, \mathcal{F}, \mathrm{P})$ the underlying probability space. The notation $\mathbb{E}_n[\cdot]$ denotes the average over index $1 \leqslant i \leqslant n$, i.e., it simply abbreviates the notation $n^{-1}\sum_{i=1}^n[\cdot]$. For example, $\mathbb{E}_n[x_{ij}^2] = n^{-1}\sum_{i=1}^n x_{ij}^2$. Moreover, we use the notation $\bar{\mathrm{E}}[\cdot] = \mathbb{E}_n[\mathrm{E}[\cdot]]$. For example, $\bar{\mathrm{E}}[v_i^2] = n^{-1}\sum_{i=1}^n \mathrm{E}[v_i^2]$. For a function $f : \mathbb{R} \times \mathbb{R} \times \mathbb{R}^p \to \mathbb{R}$, we write $\mathbb{G}_n(f) = n^{-1/2}\sum_{i=1}^n (f(y_i, d_i, x_i) - \mathrm{E}[f(y_i, d_i, x_i)])$. The l_2-norm is denoted by $\|\cdot\|$, and the l_0-norm, $\|\cdot\|_0$, denotes the number of non-zero components of a vector. Denote by $\|\cdot\|_\infty$ the maximal absolute element of a vector. For a sequence $(z_i)_{i=1}^n$ of constants, we write $\|z_i\|_{2,n} = \sqrt{\mathbb{E}_n[z_i^2]}$. For example, for a vector $\delta \in \mathbb{R}^p$, $\|x_i'\delta\|_{2,n} = \sqrt{\mathbb{E}_n[(x_i'\delta)^2]}$ denotes the prediction norm of δ. Given a vector $\delta \in \mathbb{R}^p$, and a set of indices $T \subset \{1, \ldots, p\}$, we denote by $\delta_T \in \mathbb{R}^p$ the vector such that $(\delta_T)_j = \delta_j$ if $j \in T$ and $(\delta_T)_j = 0$ if $j \notin T$. Also we write the support of δ as $\text{support}(\delta) = \{j \in \{1, ..., p\} : \delta_j \neq 0\}$. We use the notation $(a)_+ = \max\{a, 0\}$, $a \vee b = \max\{a, b\}$, and $a \wedge b = \min\{a, b\}$. We also use the notation $a \lesssim b$ to denote $a \leqslant cb$ for some constant $c > 0$ that does not depend on n; and $a \lesssim_P b$ to denote $a = O_P(b)$. The arrow $\rightsquigarrow$ denotes convergence in distribution.

We assume that the quantities such as p (the dimension of x_i), s (a bound on the numbers of non-zero elements of β_0 and θ_0), and hence $y_i, x_i, \beta_0, \theta_0, T$ and T_d are all dependent on the sample size n, and allow for the case where $p = p_n \to \infty$ and $s = s_n \to \infty$ as $n \to \infty$. However, for the notational convenience, we shall omit the dependence of these quantities on n.

2. The Methods, Conditions, and Results

2.1. **The methods.** Each of the steps outlined before uses a different identification condition. Several combinations are possible to implement each step, two of which are the following.

Algorithm 1 (Based on Post-Model Selection estimators).

(1) Run Post-ℓ_1-penalized LAD (1.3) of y_i on d_i and x_i; keep fitted value $x_i'\widetilde{\beta}$.
(2) Run Post-Lasso (1.7) of d_i on x_i; keep the residual $\widehat{v}_i := d_i - x_i'\widetilde{\theta}$.
(3) Run Instrumental LAD regression (1.8) of $y_i - x_i'\widetilde{\beta}$ on d_i using $\widehat{v}_i$ as the instrument for d_i to compute the estimator $\check{\alpha}$. Report $\check{\alpha}$ and/or perform inference based upon (1.9) or (1.12).

Algorithm 2 (Based on Regularized Estimators).

(1) Run ℓ_1-penalized LAD (1.2) of y_i on d_i and x_i; keep fitted value $x_i'\widehat{\beta}$.
(2) Run Lasso of (1.6) d_i on x_i; keep the residual $\widehat{v}_i := d_i - x_i'\widehat{\theta}$.
(3) Run Instrumental LAD regression (1.8) of $y_i - x_i'\widehat{\beta}$ on d_i using $\widehat{v}_i$ as the instrument for d_i to compute the estimator $\check{\alpha}$. Report $\check{\alpha}$ and/or perform inference based upon (1.9) or (1.12).

Comment 2.1 (Penalty Levels)**.** In order to perform ℓ_1-LAD and Lasso, one has to suitably choose the penalty levels. In the Supplementary Appendix D we provide implementation details including penalty choices for each step of the algorithm, and in all what follows we shall obey the penalty choices described in Appendix D.

Comment 2.2 (Differences)**.** Algorithm 1 relies on Post-ℓ_1-LAD and Post-Lasso while Algorithm 2 relies on ℓ_1-LAD and Lasso. Since Algorithm 1 refits the non-zero coefficients without the penalty term it has a smaller bias. Therefore it does rely on ℓ_1-LAD and Lasso obtaining sparse solutions which in turn

typically relies on restricted isometry conditions [3, 2]. Algorithm 2 relies on penalized estimators. Step 3 of both algorithms relies on instrumental LAD regression with estimated data.

Comment 2.3 (Alternative Implementations). As discussed before, the three step approach proposed here can be implemented with several different methods each with specific features. For instance, Dantzig selector, square-root Lasso or the associated post-model selection could be used instead of Lasso or Post-Lasso. Moreover, the instrumental LAD regression can be substituted by a 1-step estimator from the ℓ_1-LAD estimator $\widehat{\alpha}$ of the form $\check{\alpha} = \widehat{\alpha} + (\mathbb{E}_n[f_\epsilon \widehat{v}_i^2])^{-1}\mathbb{E}_n[\varphi(y_i - d_i\widehat{\alpha} - x_i'\widehat{\beta})\widehat{v}_i]$ or by a LAD regression with all the covariates selected in Steps 1 and 2.

2.2. **Regularity Conditions.** Here we provide regularity conditions that are sufficient for validity of the main estimation and inference results. We begin by stating our main condition, which contains the previously defined approximate sparsity as well as other more technical assumptions. Throughout the paper, let c and C be positive constants independent of n, and let $\ell_n \nearrow \infty, \delta_n \searrow 0$, and $\Delta_n \searrow 0$ be sequences of positive constants. Let $K_x := \max_{1\leqslant i \leqslant n} \|x_i\|_\infty$.

Condition I. *(i) $(\epsilon_i)_{i=1}^n$ is a sequence of i.i.d. random variables with common distribution function F such that $F(0) = 1/2$, $(v_i)_{i=1}^n$ is a sequence of independent mean-zero random variables independent of $(\epsilon_i)_{i=1}^n$, and $(x_i)_{i=1}^n$ is a sequence of non-stochastic vectors in $\mathbb{R}^p$ of covariates normalized in such a way that $\mathbb{E}_n[x_{ij}^2] = 1$ for all $1 \leqslant j \leqslant p$. The sequence $\{(y_i, d_i)'\}_{i=1}^n$ of random vectors are generated according to models (1.1) and (1.4). (ii) $c \leqslant \mathrm{E}[v_i^2] \leqslant C$ for all $1 \leqslant i \leqslant n$, and $\bar{\mathrm{E}}[d_i^4] + \bar{\mathrm{E}}[v_i^4] + \max_{1\leqslant j\leqslant p}(\bar{\mathrm{E}}[x_{ij}^2 d_i^2] + \bar{\mathrm{E}}[|x_{ij}v_i|^3]) \leqslant C$. (iii) There exists $s = s_n \geqslant 1$ such that $\|\beta_0\|_0 \leqslant s$ and $\|\theta_0\|_0 \leqslant s$. (iv) The error distribution F is absolutely continuous with continuously differentiable density $f_\epsilon(\cdot)$ such that $f_\epsilon(0) \geqslant c > 0$ and $f_\epsilon(t) \vee |f_\epsilon'(t)| \leqslant C$ for all $t \in \mathbb{R}$, and (v) $(K_x^4 + K_x^2 s^2 + s^3)\log^3(p \vee n) \leqslant n\delta_n$.*

Comment 2.4. Condition I(i) imposes the setting discussed in the previous section with the zero conditional median of the error distribution. Condition I(ii) imposes moment conditions on the structural errors and regressors to ensure good model selection performance of Lasso applied to equation (1.4). The approximate sparsity I(iii) imposes sparsity of the high-dimensional vectors β_0 and θ_0. In the theorems below we provide the required technical conditions on the growth of $s\log p$ since it is dependent on the choice of algorithm. Condition I(iv) is a set of standard assumptions in the LAD literature (see [14]) and in the instrumental quantile regression literature [10]. Condition I(v) restricts the sparsity index, so that $s^3\log^3(p\vee n) = o(n)$ is required; this is analogous to the standard assumption $s^3(\log n)^2 = o(n)$ (see [11]) invoked in the LAD analysis without any selection (i.e, where $p = s$). Most importantly, no assumptions on the separation from zero of the non-zero coefficients of θ_0 and β_0 are made.

The next condition concerns the behavior of the Gram matrix $\mathbb{E}_n[\tilde{x}_i\tilde{x}_i']$ where $\tilde{x}_i = (d_i, x_i')'$. Whenever $p+1 > n$, the empirical Gram matrix $\mathbb{E}_n[\tilde{x}_i\tilde{x}_i']$ does not have full rank and in principle is not well-behaved. However, we only need good behavior of smaller submatrices. Define the minimal and maximal m-sparse eigenvalue of $\mathbb{E}_n[\tilde{x}_i\tilde{x}_i']$ as

$$\phi_{\min}(m) := \min_{1\leqslant\|\delta\|_0\leqslant m} \frac{\delta'\mathbb{E}_n[\tilde{x}_i\tilde{x}_i']\delta}{\|\delta\|^2} \quad \text{and} \quad \phi_{\max}(m) := \max_{1\leqslant\|\delta\|_0\leqslant m} \frac{\delta'\mathbb{E}_n[\tilde{x}_i\tilde{x}_i']\delta}{\|\delta\|^2}. \tag{2.13}$$

To assume that $\phi_{\min}(m) > 0$ requires that all empirical Gram submatrices formed by any m components of $\tilde{x}_i$ are positive definite. We shall employ the following condition as a sufficient condition for our results.

Condition SE. *There exists a sequence of constants $\ell_n \to \infty$ such that the maximal and minimal $\ell_n s$-sparse eigenvalues are bounded from below and away from zero, namely with probability at least $1 - \Delta_n$,*

$$\kappa' \leqslant \phi_{\min}(\ell_n s) \leqslant \phi_{\max}(\ell_n s) \leqslant \kappa'',$$

where $0 < \kappa' < \kappa'' < \infty$ are constants independent of n.

Comment 2.5. Condition SE is quite plausible for many designs of interest. Essentially it can be established by combining tail conditions of the regressors and a growth restriction on s and p relative to n. For instance, Theorem 3.2 in [22] (see also [28] and [1]) shows that Condition SE holds for i.i.d. zero-mean sub-Gaussian regressors and $s \log^2(n \vee p) \leqslant \delta_n n$; while Theorem 1.8 [22] (see also Lemma 1 in [4]) shows that Condition SE holds for i.i.d. uniformly bounded zero-mean regressors and $s(\log^3 n)\log(p \vee n) \leqslant \delta_n n$.

2.3. **Results.** We begin with considering Algorithm 1.

Theorem 1 (Robust Inference, Algorithm 1)**.** *Let $\check{\alpha}$ be obtained by Algorithm 1. Suppose that Conditions I and SE are satisfied for all $n \geqslant 1$. Moreover, suppose that with probability at least $1 - \Delta_n$, $\|\widetilde{\beta}\|_0 \leqslant Cs$. Then, as $n \to \infty$ and for $\sigma_n^2 = 1/(4 f_\epsilon^2 \bar{\mathrm{E}}[v_i^2])$,*

$$\sigma_n^{-1}\sqrt{n}(\check{\alpha} - \alpha_0) \rightsquigarrow N(0,1) \text{ and } nL_n(\alpha_0) \rightsquigarrow \chi^2(1).$$

Theorem 1 establishes the first main result of the paper. Theorem 1 relies on the post model selection estimators which in turn hinge on achieving sufficiently sparse estimates $\widehat{\beta}$ and $\widehat{\theta}$. Sparsity of the former can be directly achieved under sharp penalty choices for optimal rates as discussed in the Supplementary Appendix D.2. The sparsity for the latter potentially requires heavier penalty as shown in [3]. Alternatively, sparsity for the estimator in Step 1 can also be achieved by truncating the smallest components of estimate $\widehat{\beta}$.[2]

Next we turn to the analysis of Algorithm 2 which relies on the regularized estimators instead of the post-model selection estimators. Theorem 2 below establishes that Algorithm 2 achieves the same inferential guarantees as the results in Theorem 1 for Algorithm 1.

Theorem 2 (Robust Inference, Algorithm 2)**.** *Let $\check{\alpha}$ be obtained by Algorithm 2. Suppose that Conditions I and SE are satisfied for all $n \geqslant 1$. Moreover, suppose that with probability at least $1 - \Delta_n$, $\|\widehat{\beta}\|_0 \leqslant Cs$. Then, as $n \to \infty$ and for $\sigma_n^2 = 1/(4 f_\epsilon^2 \bar{\mathrm{E}}[v_i^2])$,*

$$\sigma_n^{-1}\sqrt{n}(\check{\alpha} - \alpha_0) \rightsquigarrow N(0,1) \text{ and } nL_n(\alpha_0) \rightsquigarrow \chi^2(1).$$

Theorem 2 establishes the second main result of the paper.

An important consequence of these results is the following corollary. Here $\mathcal{Q}_n$ denotes a collection of distributions for $\{(y_i, d_i)'\}_{i=1}^n$ and for $Q_n \in \mathcal{Q}_n$ the notation P_{Q_n} means that under P_{Q_n}, $\{(y_i, d_i)'\}_{i=1}^n$ is distributed according to Q_n.

[2]Lemma 3 in Appendix C formally shows that a suitable truncation preserves the rate of convergence under our conditions.

Corollary 1 (**Uniformly Valid Confidence Intervals**). *Let $\check{\alpha}$ be the estimator of α_0 constructed according to Algorithm 1 (resp. Algorithm 2) and let $\mathcal{Q}_n$ be the collection of all distributions of $\{(y_i, d_i)'\}_{i=1}^n$ for which the conditions of Theorem 1 (resp. Theorem 2) are satisfied for given $n \geqslant 1$. Then as $n \to \infty$, uniformly in $Q_n \in \mathcal{Q}_n$*

$$\mathrm{P}_{Q_n}(\alpha_0 \in [\check{\alpha} \pm \sigma_n z_{\xi/2}/\sqrt{n}]) \to 1-\xi \quad \textit{and} \quad \mathrm{P}_{Q_n}(\alpha_0 \in \widehat{A}_{n,\xi}) \to 1-\xi,$$

where $z_{\xi/2} = \Phi^{-1}(1-\xi/2)$ and $\widehat{A}_{n,\xi} = \{\alpha \in \mathcal{A} : nL_n(\alpha) \leqslant (1-\xi)\text{-quantile of } \chi^2(1)\}$.

Corollary 1 establishes the third main result of the paper; it highlights the uniformity nature of the results. As long as the overall sparsity requirements hold, imperfect model selection in Steps 1 and 2 do not compromise the results. The robustness of the approach is also apparent from the fact that Corollary 1 allows for the data-generating process to change with n. This result is new even under the traditional case of fixed-p asymptotics. Condition I and SE together with the appropriate side conditions in the theorems explicitly characterize regions of data-generating processes for which the uniformity result holds. Simulations results discussed next also provide an additional evidence that these regions are substantial.

3. Monte-Carlo Experiments

In this section we examine the finite sample performance of the proposed estimators. We focus on the estimator associated with Algorithm 1 based on post-model selection methods.

We considered the following regression model:

$$y = d\alpha_0 + x'(c_y\theta_0) + \epsilon, \qquad d = x'(c_d\theta_0) + v, \tag{3.14}$$

where $\alpha_0 = 1/2$, $\theta_{0j} = 1/j^2, j = 1, \ldots, 10$, and $\theta_{0j} = 0$ otherwise, $x = (1, z')'$ consists of an intercept and covariates $z \sim N(0, \Sigma)$, and the errors ϵ and v are independently and identically distributed as $N(0,1)$. The dimension p of the covariates x is 300, and the sample size n is 250. The regressors are correlated with $\Sigma_{ij} = \rho^{|i-j|}$ and $\rho = 0.5$. The coefficients c_y and c_d are used to control the R^2 of the reduce form equation. For each equation, we consider the following values for the R^2: $\{0, 0.1, 0.2, \ldots, 0.8, 0.9\}$. Therefore we have 100 different designs and results are based on 500 repetitions for each design. For each repetition we draw new vectors x_i's and errors ϵ_i's and v_i's.

The design above with $x'(c_y\theta_0)$ is a sparse model. However, the decay of the components of θ_0 rules out typical "separation from zero" assumptions of the coefficients of "important" covariates (since the last component is of the order of $1/n$), unless c_y is very large. Thus, we anticipate that "standard" post-selection inference procedures – which rely on model selection of the outcome equation only – work poorly in the simulation study. In contrast, based upon the prior theoretical arguments, we anticipate that our instrumental LAD estimator– which works off both equations in (3.14)– to work well in the simulation study.

The simulation study focuses on Algorithm 1. Standard errors are computed using the formula (1.10). (Algorithm 2 worked similarly, though somewhat worse due to larger biases). As the main benchmark

we consider the standard post-model selection estimator $\widetilde{\alpha}$ based on the post ℓ_1-penalized LAD method, as defined in (1.3).

In Figure 1, we display the (empirical) rejection probability of tests of a true hypothesis $\alpha = \alpha_0$, with nominal size of tests equal to 0.05. The left-top plot shows the rejection frequency of the standard post-model selection inference procedure based upon $\widetilde{\alpha}$ (where the inference procedure assumes perfect recovery of the true model). The rejection frequency deviates very sharply from the ideal rejection frequency of 0.05. This confirms the anticipated failure (lack of uniform validity) of inference based upon the standard post-model selection procedure in designs where coefficients are not well separated from zero (so that perfect recovery does not happen). In sharp contrast, the right top and bottom plots show that both of our proposed procedures (based on estimator $\check{\alpha}$ and the result (1.9) and on the statistic L_n and the result (1.12)) perform well, closely tracking the ideal level of 0.05. This is achieved uniformly over all the designs considered in the study, and this confirms our theoretical results established in Corollary 1.

In Figure 2, we compare the performance of the standard post-selection estimator $\widetilde{\alpha}$ (defined in (1.3)) and our proposed post-selection estimator $\check{\alpha}$ (obtained via Algorithm 1) . We display results in three different metrics of performance – mean bias (top row), standard deviation (middle row), and root mean square error (bottom row) of the two approaches. The significant bias for the standard post-selection procedure occurs when the indirect equation (1.4) is nontrivial, that is, when the main regressor is correlated to other controls. Such bias can be positive or negative depending on the particular design. The proposed post-selection estimator $\check{\alpha}$ performs well in all three metrics. The root mean square error for the proposed estimator $\check{\alpha}$ are typically much smaller than those for standard post-model selection estimators $\widetilde{\alpha}$ (as shown by bottom plots in Figure 2). This is fully consistent with our theoretical results and minimax efficiency considerations given in Section 5.

4. Generalization to Heteroscedastic Case

We emphasize that both proposed algorithms exploit the homoscedasticity of the model (1.1) with respect to the error term ϵ_i. The generalization to the heteroscedastic case can be achieved as follows. In order to achieve the semiparametric efficiency bound we need to consider the weighted version of the auxiliary equation (1.4). Specifically, we can rely on the following of weighted decomposition:

$$f_i d_i = f_i x_i' \theta_0^* + v_i^*, \quad \mathrm{E}[f_i v_i^*] = 0, \quad i = 1, \ldots, n, \tag{4.15}$$

where the weights are conditional densities of error terms ϵ_i evaluated at their medians of 0,

$$f_i = f_{\epsilon_i}(0|d_i, x_i), \quad i = 1, \ldots, n, \tag{4.16}$$

which in general vary under heteroscedasticity. With that in mind it is straightforward to adapt the proposed algorithms when the weights $(f_i)_{i=1}^n$ are known. For example Algorithm 1 becomes as follows.

Algorithm 1′ (Based on Post-Model Selection estimators).

(1) Run Post-ℓ_1-penalized LAD of y_i on d_i and x_i; keep fitted value $x_i'\widetilde{\beta}$.
(2) Run Post-Lasso of $f_i d_i$ on $f_i x_i$; keep the residual $\widehat{v}_i^* := f_i(d_i - x_i'\widetilde{\theta})$.

(3) Run Instrumental LAD regression of $y_i - x_i'\widetilde{\beta}$ on d_i using $\widehat{v}_i^*$ as the instrument for d_i to compute the estimator $\check{\alpha}$. Report $\check{\alpha}$ and/or perform inference.

An analogous generalization of Algorithm 2 based on regularized estimator results from removing the word "Post" in the algorithm above.

Under similar regularity conditions, uniformly over a large collection $\mathcal{Q}_n^*$ of distributions of $\{(y_i, d_i)'\}_{i=1}^n$, the estimator $\check{\alpha}$ above obeys

$$(4\bar{\mathrm{E}}[v_i^{*2}])^{1/2}\sqrt{n}(\check{\alpha} - \alpha_0) \rightsquigarrow N(0,1). \tag{4.17}$$

Moreover, the criterion function at the true value α_0 in Step 3 also has a pivotal behavior, namely

$$nL_n(\alpha_0) \rightsquigarrow \chi^2(1), \tag{4.18}$$

which can also be used to construct a confidence region $\widehat{A}_{n,\xi}$ based on the L_n-statistic as in (1.12) with coverage $1-\xi$ uniformly over the collection of distributions $\mathcal{Q}_n^*$.

In practice the density function values $(f_i)_{i=1}^n$ are typically unknown and need to be replaced by estimates $(\widehat{f}_i)_{i=1}^n$. The analysis of the impact of such estimation is very delicate and is developed in the companion work [8], which considers the more general problem of uniformly valid inference for quantile regression models in approximately sparse models.

5. Discussion and Conclusion

5.1. **Connection to Neymanization.** In this section we make some connections to Neyman's $C(\alpha)$ test ([19, 20]). For the sake of exposition we assume that $(y_i, x_i, d_i)_{i=1}^n$ are i.i.d. but we shall use the heteroscedastic setup introduced in the previous section. We consider the estimating equation for α_0:

$$\mathrm{E}[\varphi(y_i - d_i\alpha_0 - x_i'\beta_0)v_i] = 0.$$

Our problem is to find useful instruments v_i such that

$$\frac{\partial}{\partial\beta}\mathrm{E}[\varphi(y_i - d_i\alpha_0 - x_i'\beta)v_i]|_{\beta=\beta_0} = 0.$$

If this property holds, the estimator of α_0 will be "immunized" against "crude" or nonregular estimation of β_0, for example, via a post-selection procedure or some regularization procedure. Such immunization ideas are in fact behind Neyman's classical construction of his $C(\alpha)$ test, so we shall use the term "Neymanization" to describe such procedure. There will be many instruments v_i that can achieve the property stated above, and there will be one that is optimal.

The instruments can be constructed by taking $v_i := z_i/f_i$, where z_i is the residual in the regression equation:

$$w_i d_i = w_i m_0(x_i) + z_i, \quad \mathrm{E}[w_i z_i | x_i] = 0, \tag{5.19}$$

where w_i is a nonnegative weight, a function of (d_i, z_i) only, for example $w_i = 1$ or $w_i = f_i$ – the latter choice will in fact be optimal. Note that function $m_0(x_i)$ solves the least squares problem

$$\min_{h\in\mathcal{H}} \mathrm{E}\left[\{w_i d_i - w_i h(x_i)\}^2\right], \tag{5.20}$$

where $\mathcal{H}$ is the class of measurable functions $h(x_i)$ such that $\mathrm{E}[w_i^2 h^2(x_i)] < \infty$. Our assumption is that the $m_0(x_i)$ is a sparse function $x_i'\theta_0$, with $\|\theta_0\|_0 \leqslant s$ so that

$$w_i d_i = w_i x_i'\theta_0 + z_i, \quad \mathrm{E}[w_i z_i | x_i] = 0. \tag{5.21}$$

In finite samples, the sparsity assumption allows to employ post-Lasso and Lasso to solve the least squares problem above approximately, and estimate z_i. Of course, the use of other structured assumptions may motivate the use of other regularization methods.

Arguments similar to those in the proofs show that, for $\sqrt{n}(\alpha - \alpha_0) = O(1)$,

$$\sqrt{n}\{\mathbb{E}_n[\varphi(y_i - d_i\alpha - x_i'\widehat{\beta})v_i] - \mathbb{E}_n[\varphi(y_i - d_i\alpha - x_i'\beta_0)v_i]\} = o_P(1),$$

for $\widehat{\beta}$ based on a sparse estimation procedure, despite the fact that $\widehat{\beta}$ converges to β_0 at a slower rate than $1/\sqrt{n}$. That is, the empirical estimating equations behave as if β_0 is known. Hence for estimation we can use $\widehat{\alpha}$ as a minimizer of the statistic:

$$L_n(\alpha) = c_n^{-1}|\sqrt{n}\mathbb{E}_n[\varphi(y_i - d_i\alpha - x_i'\widehat{\beta})v_i]|^2,$$

where $c_n = \mathbb{E}_n[v_i^2]/4$. Since $L_n(\alpha_0) \rightsquigarrow \chi^2(1)$, we can also use the statistic directly for testing hypotheses and for construction of confidence sets.

This is in fact a version of Neyman's $C(\alpha)$ test statistic, adapted to the present non-smooth setting. The usual expression of $C(\alpha)$ statistic is different. To see a more familiar form, note that $\theta_0 = \mathrm{E}[w_i^2 x_i x_i']^{-}\mathrm{E}[w_i^2 d_i x_i']$, where A^- denotes a generalized inverse of A, and write

$$v_i = (w_i/f_i)d_i - (w_i/f_i)x_i'\mathrm{E}[w_i^2 x_i x_i']^{-}\mathrm{E}[w_i^2 d_i x_i'], \quad \text{and} \quad \widehat{\varphi}_i := \varphi(y_i - d_i\alpha - x_i'\widehat{\beta}),$$

so that,

$$L_n(\alpha) = c_n^{-1}|\sqrt{n}\{\mathbb{E}_n[\widehat{\varphi}_i(w_i/f_i)d_i] - \mathbb{E}_n[\widehat{\varphi}_i(w_i/f_i)x_i]'\mathrm{E}[w_i^2 x_i x_i']^{-}\mathrm{E}[w_i^2 d_i x_i']\}|^2.$$

This is indeed a familiar form of a $C(\alpha)$ statistic.

The estimator $\widehat{\alpha}$ that minimizes L_n up to $o_P(1)$, under suitable regularity conditions,

$$\sigma_n^{-1}\sqrt{n}(\widehat{\alpha} - \alpha_0) \rightsquigarrow N(0,1), \quad \sigma_n^2 = \frac{1}{4}\mathrm{E}[f_i d_i v_i]^{-2}\mathrm{E}[v_i^2].$$

The smallest value of σ_n^2 is achieved by using $v_i = v_i^*$ induced by setting $w_i = f_i$:

$$\sigma_n^{*2} = \frac{1}{4}\mathrm{E}[v_i^{*2}]^{-1}. \tag{5.22}$$

Thus, setting $w_i = f_i$ gives an optimal instrument v_i^* amongst all "immunizing" instruments generated by the process described above. Obviously, this improvement translates into shorter confidence intervals and better testing based on either $\widehat{\alpha}$ or L_n. While $w_i = f_i$ is optimal, f_i will have to be estimated in practice, resulting actually in more stringent condition than when using non-optimal, known weights, e.g., $w_i = 1$. The use of known weights may also give better behavior under misspecification of the model. Under homoscedasticity, $w_i = 1$ is an optimal weight.

5.2. **Minimax Efficiency.** There is also a clean connection to the (local) minimax efficiency analysis from the semiparametric efficiency analysis. [16] derives an efficient score function for the partially linear median regression model:

$$S_i = 2\varphi(y_i - d_i\alpha_0 - x_i'\beta_0)f_i[d_i - m_0^*(x)],$$

where $m_0^*(x_i)$ is $m_0(x_i)$ in (5.19) induced by the weight $w_i = f_i$:

$$m_0^*(x_i) = \frac{\mathrm{E}[f_i^2 d_i|x_i]}{\mathrm{E}[f_i^2|x_i]}.$$

Using the assumption $m_0^*(x_i) = x_i'\theta_0^*$, where $\|\theta_0^*\|_0 \leqslant s \ll n$ is sparse, we have that

$$S_i = 2\varphi(y_i - d_i\alpha_0 - x_i'\beta_0)v_i^*,$$

which is the score that was constructed using Neymanization. It follows that the estimator based on the instrument v_i^* is actually efficient in the minimax sense (see Theorem 18.4 in [15]), and inference about α_0 based on this estimator provides best minimax power against local alternatives (see Theorem 18.12 in [15]).

The claim above is formal as long as, given a law Q_n^*, the least favorable submodels are permitted as deviations that lie within the overall model $\mathcal{Q}_n$. Specifically, given a law Q_n^*, we shall need to allow for a certain neighborhood $\mathcal{Q}_n^\delta$ of Q_n^* such that $Q_n^* \in \mathcal{Q}_n^\delta \subset \mathcal{Q}_n$, where the overall model $\mathcal{Q}_n$ is defined similarly as before, except now permitting heteroscedasticity (or we can keep homoscedasticity $f_i = f_\epsilon$ to maintain formality). To allow for this we consider a collection of laws indexed by a parameter $t = (t_1, t_2)$, generated by:

$$y_i = d_i \underbrace{(\alpha_0^* + t_1)}_{\alpha_0} + x_i' \underbrace{(\beta_0^* + t_2\theta_0^*)}_{\beta_0} + \epsilon_i, \quad \|t\| \leqslant \delta, \tag{5.23}$$

$$f_i d_i = f_i x_i'\theta_0^* + v_i^*, \quad \mathrm{E}[f_i v_i^*|x_i] = 0, \tag{5.24}$$

where $\|\beta_0^*\|_0 + \|\theta_0^*\|_0 \leqslant s$ and conditions as in Section 2 hold. The case with $t = 0$ generates the law Q_n^*; by varying t within δ-ball, we generate the set of laws, denoted $\mathcal{Q}_n^\delta$, containing the least favorable deviations from $t = 0$. By [16], the efficient score for the model given above is S_i, so we cannot have a better regular estimator than the estimator whose influence function is $J^{-1}S_i$, where $J = \mathrm{E}[S_i^2]$. Since our overall model $\mathcal{Q}_n$ contains $\mathcal{Q}_n^\delta$, all the formal conclusions about (local minimax) optimality of our estimators hold from theorems cited above (using subsequence arguments to handle models changing with n). Our estimators are regular, since under any law Q_n in the set $\mathcal{Q}_n^\delta$ with $\delta \to 0$, the first order asymptotics of $\sqrt{n}(\check{\alpha} - \alpha_0)$ does not change, as a consequence of theorems in Section 2 (in fact our theorems show more than this).

5.3. **Conclusion.** In this paper we propose a method for inference on the coefficient α_0 of a main regressor that holds uniformly over many data-generating process which is robust to possible "moderate" model selection mistakes. The robustness of the method is achieved by relying on a Neyman type estimating equation whose gradient with respect to the nuisance parameters is zero. In the present homoscedastic setting the proposed estimator is asymptotically normal and also achieves the semi-parametric efficiency bound.

Appendix A. Instrumental LAD Regression with Estimated Inputs

Throughout this section, let

$$\begin{aligned}\psi_{\alpha,\beta,\theta}(y_i,d_i,x_i) &= (1/2 - 1\{y_i \leqslant x_i'\beta + d_i\alpha\})(d_i - x_i'\theta)\\ &= (1/2 - 1\{y_i \leqslant x_i'\beta + d_i\alpha\})\{v_i - x_i'(\theta-\theta_0)\}.\end{aligned}$$

For fixed $\alpha \in \mathbb{R}$ and $\beta,\theta \in \mathbb{R}^p$, define the function

$$\Gamma(\alpha,\beta,\theta) := \bar{\mathrm{E}}[\psi_{\alpha,\beta,\theta}(y_i,d_i,x_i)].$$

For the notational convenience, let $h = (\beta',\theta')'$, $h_0 = (\beta_0',\theta_0')'$ and $\widehat{h} = (\widehat{\beta}',\widehat{\theta}')'$. The partial derivative of $\Gamma(\alpha,\beta,\theta)$ with respect to α is denoted by $\Gamma_1(\alpha,\beta,\theta)$ and the partial derivative of $\Gamma(\alpha,\beta,\theta)$ with respect to $h = (\beta',\theta')'$ is denoted by $\Gamma_2(\alpha,\beta,\theta)$. Consider the following high-level condition. Here $(\widehat{\beta}',\widehat{\theta}')'$ is a generic estimator of $(\beta_0',\theta_0')'$ (and not necessarily ℓ_1-LAD and Lasso estimators, reps.), and $\check{\alpha}$ is defined by $\check{\alpha} \in \arg\min_{\alpha\in\mathcal{A}} L_n(\alpha)$ with this $(\widehat{\beta}',\widehat{\theta}')'$, where $\mathcal{A}$ here is also a generic (possibly random) compact interval. We assume that $(\widehat{\beta}',\widehat{\theta}')',\mathcal{A}$ and $\check{\alpha}$ satisfy the following conditions.

Condition ILAD. (i) $f_\epsilon(t) \vee |f_\epsilon'(t)| \leqslant C$ for all $t \in \mathbb{R}$, $\bar{\mathrm{E}}[v_i^2] \geqslant c > 0$ and $\bar{\mathrm{E}}[v_i^4] \vee \bar{\mathrm{E}}[d_i^4] \leqslant C$. Moreover, for some sequences $\delta_n \searrow 0$ and $\Delta_n \searrow 0$, with probability at least $1-\Delta_n$,
(ii) $\{\alpha : |\alpha-\alpha_0| \leqslant n^{-1/2}/\delta_n\} \subset \mathcal{A}$, where $\mathcal{A}$ is a (possibly random) compact interval;
(iii) the estimated parameters $(\widehat{\beta}',\widehat{\theta}')'$ satisfy

$$\big\{1 \vee \max_{1\leqslant i\leqslant n}(\mathrm{E}[|v_i|] \vee |x_i'(\widehat{\theta}-\theta_0)|)\big\}^{1/2}\|x_i'(\widehat{\beta}-\beta_0)\|_{2,n} \leqslant \delta_n n^{-1/4},\ \ \|x_i'(\widehat{\theta}-\theta_0)\|_{2,n} \leqslant \delta_n n^{-1/4}, \tag{A.25}$$

$$\sup_{\alpha\in\mathcal{A}} |\mathbb{G}_n(\psi_{\alpha,\widehat{\beta},\widehat{\theta}} - \psi_{\alpha,\beta_0,\theta_0})| \leqslant \delta_n, \tag{A.26}$$

where recall that $\mathbb{G}_n(f) = n^{-1/2}\sum_{i=1}^n \{f(y_i,d_i,x_i) - \mathrm{E}[f(y_i,d_i,x_i)]\}$; and lastly
(iv) the estimator $\check{\alpha}$ satisfies $|\check{\alpha}-\alpha_0| \leqslant \delta_n$.

Comment A.1. Condition ILAD suffices to make the impact of the estimation of instruments negligible on the first order asymptotics of the estimator $\check{\alpha}$. We note that Condition ILAD covers several different estimators including both estimators proposed in Algorithms 1 and 2.

The following lemma summarizes the main inferential result based on the high level Condition ILAD.

Lemma 1. *Under Condition ILAD we have, for $\sigma_n^2 = 1/(4f_\epsilon^2\bar{\mathrm{E}}[v_i^2])$,*

$$\sigma_n^{-1}\sqrt{n}(\check{\alpha}-\alpha_0) \rightsquigarrow N(0,1) \ \textit{and}\ \ nL_n(\alpha_0) \rightsquigarrow \chi^2(1),$$

Proof of Lemma 1. We shall separate the proof into two parts.

Part 1. (Proof for the first assertion). Observe that

$$\begin{aligned}\mathbb{E}_n[\psi_{\check{\alpha},\widehat{\beta},\widehat{\theta}}(y_i,d_i,x_i)] &= \mathbb{E}_n[\psi_{\alpha_0,\beta_0,\theta_0}(y_i,d_i,x_i)] + \mathbb{E}_n[\psi_{\check{\alpha},\widehat{\beta},\widehat{\theta}}(y_i,d_i,x_i) - \psi_{\alpha_0,\beta_0,\theta_0}(y_i,d_i,x_i)]\\ &= \mathbb{E}_n[\psi_{\alpha_0,\beta_0,\theta_0}(y_i,d_i,x_i)] + \Gamma(\check{\alpha},\widehat{\beta},\widehat{\theta})\\ &+ n^{-1/2}\mathbb{G}_n(\psi_{\check{\alpha},\widehat{\beta},\widehat{\theta}} - \psi_{\check{\alpha},\beta_0,\theta_0}) + n^{-1/2}\mathbb{G}_n(\psi_{\check{\alpha},\beta_0,\theta_0} - \psi_{\alpha_0,\beta_0,\theta_0})\\ &= I + II + III + IV.\end{aligned}$$

By Condition ILAD(iii) (A.26) we have with probability at least $1-\Delta_n$ that $|III| \leqslant \delta_n n^{-1/2}$. We wish to show that

$$\left|II + (f_\epsilon \bar{\mathrm{E}}[v_i^2])(\check{\alpha}-\alpha_0)\right| \lesssim_P \delta_n n^{-1/2} + \delta_n |\check{\alpha}-\alpha_0|. \tag{A.27}$$

Observe that

$$\begin{aligned}\Gamma(\alpha,\widehat{\beta},\widehat{\theta}) &= \Gamma(\alpha,\beta_0,\theta_0) + \Gamma(\alpha,\widehat{\beta},\widehat{\theta}) - \Gamma(\alpha,\beta_0,\theta_0)\\ &= \Gamma(\alpha,\beta_0,\theta_0) + \{\Gamma(\alpha,\widehat{\beta},\widehat{\theta}) - \Gamma(\alpha,\beta_0,\theta_0) - \Gamma_2(\alpha,\beta_0,\theta_0)'(\widehat{h}-h_0)\} + \Gamma_2(\alpha,\beta_0,\theta_0)'(\widehat{h}-h_0).\end{aligned}$$

Since $\Gamma(\alpha_0,\beta_0,\theta_0)=0$, by Taylor's theorem, there exists some point $\tilde{\alpha}$ between α_0 and α such that $\Gamma(\alpha,\beta_0,\theta_0) = \Gamma_1(\tilde{\alpha},\beta_0,\theta_0)(\alpha-\alpha_0)$. By its definition, we have

$$\Gamma_1(\alpha,\beta,\theta) = -\bar{\mathrm{E}}[f_\epsilon(x_i'(\beta-\beta_0)+d_i(\alpha-\alpha_0))d_i(d_i-x_i'\theta)] = -\bar{\mathrm{E}}[f_\epsilon(x_i'(\beta-\beta_0)+d_i(\alpha-\alpha_0))d_i\{v_i-x_i'(\theta-\theta_0)\}].$$

Since $f_\epsilon = f_\epsilon(0)$ and $d_i = x_i'\theta_0 + v_i$ with $\mathrm{E}[v_i]=0$, we have $\Gamma_1(\alpha_0,\beta_0,\theta_0) = -f_\epsilon\bar{\mathrm{E}}[d_iv_i] = -f_\epsilon\bar{\mathrm{E}}[v_i^2]$. Also

$$|\Gamma_1(\alpha,\beta_0,\theta_0) - \Gamma_1(\alpha_0,\beta_0,\theta_0)| \leqslant |\bar{\mathrm{E}}[\{f_\epsilon(0) - f_\epsilon(d_i(\alpha-\alpha_0))d_iv_i]| \leqslant C|\alpha-\alpha_0|\bar{\mathrm{E}}[|d_i^2v_i|].$$

Hence $\Gamma(\check{\alpha},\beta_0,\theta_0) = -f_\epsilon\bar{\mathrm{E}}[v_i^2] + O(1)|\check{\alpha}-\alpha_0|$.

Observe that

$$\Gamma_2(\alpha,\beta,\theta) = \begin{pmatrix} -\bar{\mathrm{E}}[f_\epsilon(x_i'(\beta-\beta_0)+d_i(\alpha-\alpha_0))(d_i-x_i'\theta)x_i] \\ -\bar{\mathrm{E}}[(1/2 - 1\{y_i \leqslant x_i'\beta + d_i\alpha\})x_i] \end{pmatrix}.$$

Note that since $\bar{\mathrm{E}}[f_\epsilon(0)(d_i - x_i'\theta_0)x_i] = f_\epsilon\bar{\mathrm{E}}[v_ix_i] = 0$ and $\bar{\mathrm{E}}[(1/2 - 1\{y_i \leqslant x_i'\beta_0 + d_i\alpha_0\})x_i] = \bar{\mathrm{E}}[(1/2-1\{\epsilon_i \leqslant 0\})x_i] = 0$, we have $\Gamma_2(\alpha_0,\beta_0,\theta_0)=0$. Moreover,

$$\begin{aligned}|\Gamma_2(\alpha,\beta_0,\theta_0)'(\widehat{h}-h_0)| &= |\{\Gamma_2(\alpha,\beta_0,\theta_0) - \Gamma_2(\alpha_0,\beta_0,\theta_0)\}'(\widehat{h}-h_0)|\\ &\leqslant |\bar{\mathrm{E}}[\{f_\epsilon(d_i(\alpha-\alpha_0)) - f_\epsilon(0)\}v_ix_i'](\widehat{\beta}-\beta_0)|\\ &\quad + |\bar{\mathrm{E}}[\{F(d_i(\alpha-\alpha_0)) - F(0)\}x_i'](\widehat{\theta}-\theta_0)]|\\ &\leqslant O(1)\{\|x_i'(\widehat{\beta}-\beta_0)\|_{2,n} + \|x_i'(\widehat{\theta}-\theta_0)\|_{2,n}\}|\alpha-\alpha_0|\\ &= O_P(\delta_n)|\alpha-\alpha_0|.\end{aligned}$$

Hence $|\Gamma_2(\check{\alpha},\beta_0,\theta_0)'(\widehat{h}-h_0)| \lesssim_P \delta_n|\check{\alpha}-\alpha_0|$.

Denote by $\Gamma_{22}(\alpha,\beta,\theta)$ the Hessian matrix of $\Gamma(\alpha,\beta,\theta)$ with respect to $h=(\beta',\theta')'$. Then

$$\Gamma_{22}(\alpha,\beta,\theta) = \begin{pmatrix} -\bar{\mathrm{E}}[f_\epsilon'(x_i'(\beta-\beta_0)+d_i(\alpha-\alpha_0))(d_i-x_i'\theta)x_ix_i'] & \bar{\mathrm{E}}[f_\epsilon(x_i'(\beta-\beta_0)+d_i(\alpha-\alpha_0))x_ix_i'] \\ \bar{\mathrm{E}}[f_\epsilon(x_i'(\beta-\beta_0)+d_i(\alpha-\alpha_0))x_ix_i'] & 0 \end{pmatrix},$$

so that

$$\begin{aligned}(\widehat{h}-h_0)'\Gamma_{22}(\alpha,\beta,\theta)(\widehat{h}-h_0) &\leqslant |(\widehat{\beta}-\beta_0)'\bar{\mathrm{E}}[f_\epsilon'(x_i'(\beta-\beta_0)+d_i(\alpha-\alpha_0))(d_i-x_i'\theta)x_ix_i'](\widehat{\beta}-\beta_0)|\\ &\quad + 2|(\widehat{\beta}-\beta_0)'\bar{\mathrm{E}}[f_\epsilon(x_i'(\beta-\beta_0)+d_i(\alpha-\alpha_0))x_ix_i'](\widehat{\theta}-\theta_0)|\\ &\leqslant C\{\max_{1\leqslant i\leqslant n}\mathrm{E}[|d_i-x_i'\theta|]\|x_i'(\widehat{\beta}-\beta_0)\|_{2,n}^2 + 2\|x_i'(\widehat{\beta}-\beta_0)\|_{2,n}\cdot\|x_i'(\widehat{\theta}-\theta_0)\|_{2,n}\}.\end{aligned}$$

Here $|d_i - x_i'\theta| = |v_i - x_i'(\theta-\theta_0)| \leqslant |v_i| + |x_i'(\theta-\theta_0)|$. Hence by Taylor's theorem together with ILAD(iii), we conclude that

$$|\Gamma(\check\alpha, \widehat\beta, \widehat\theta) - \Gamma(\check\alpha, \beta_0, \theta_0) - \Gamma_2(\check\alpha, \beta_0, \theta_0)'(\widehat h - h_0)| \lesssim_P \delta_n n^{-1/2}.$$

This leads to the expansion in (A.27).

We now proceed to bound the fourth term. By Condition ILAD(iii) we have with probability at least $1-\Delta_n$ that $|\check\alpha - \alpha_0| \leqslant \delta_n$. Observe that

$$\begin{aligned}(\psi_{\alpha,\beta_0,\theta_0} - \psi_{\alpha_0,\beta_0,\theta_0})(y_i, d_i, x_i) &= (1\{y_i \leqslant x_i'\beta_0 + d_i\alpha_0\} - 1\{y_i \leqslant x_i'\beta_0 + d_i\alpha\})v_i \\ &= (1\{\epsilon_i \leqslant 0\} - 1\{\epsilon_i \leqslant d_i(\alpha-\alpha_0)\})v_i,\end{aligned}$$

so that $|(\psi_{\alpha,\beta_0,\theta_0} - \psi_{\alpha_0,\beta_0,\theta_0})(y_i, d_i, x_i)| \leqslant 1\{|\epsilon_i| \leqslant \delta_n|d_i|\}|v_i|$ whenever $|\alpha-\alpha_0| \leqslant \delta_n$. Since the class of functions $\{(y,d,x) \mapsto (\psi_{\alpha,\beta_0,\theta_0} - \psi_{\alpha_0,\beta_0,\theta_0})(y,d,x) : |\alpha-\alpha_0| \leqslant \delta_n\}$ is a VC subgraph class with VC index bounded by some constant independent of n, using (a version of) Theorem 2.14.1 in [25], we have

$$\sup_{|\alpha-\alpha_0|\leqslant\delta_n} |\mathbb{G}_n(\psi_{\alpha,\beta_0,\theta_0} - \psi_{\alpha_0,\beta_0,\theta_0})| \lesssim_P (\bar{\mathrm{E}}[1\{|\epsilon_i| \leqslant \delta_n|d_i|\}v_i^2])^{1/2} \lesssim_P \delta_n^{1/2}.$$

This implies that $|IV| \lesssim_P \delta_n^{1/2} n^{-1/2}$.

Combining these bounds on II, III and IV, we have the following stochastic expansion

$$\mathbb{E}_n[\psi_{\check\alpha,\widehat\beta,\widehat\theta}(y_i,d_i,x_i)] = -(f_\epsilon \bar{\mathrm{E}}[v_i^2])(\check\alpha-\alpha_0) + \mathbb{E}_n[\psi_{\alpha_0,\beta_0,\theta_0}(y_i,d_i,x_i)] + O_P(\delta_n^{1/2}n^{-1/2}) + O_P(\delta_n)|\check\alpha-\alpha_0|.$$

Let $\alpha^* = \alpha_0 + (f_\epsilon\bar{\mathrm{E}}[v_i^2])^{-1}\mathbb{E}_n[\psi_{\alpha_0,\beta_0,\theta_0}(y_i,d_i,x_i)]$. Then $\alpha^* \in \mathcal{A}$ with probability $1-o(1)$ since $|\alpha^*-\alpha_0| \lesssim_P n^{-1/2}$. It is not difficult to see that the above stochastic expansion holds with $\check\alpha$ replaced by α^*, so that

$$\mathbb{E}_n[\psi_{\alpha^*,\widehat\beta,\widehat\theta}(y_i,d_i,x_i)] = -(f_\epsilon \bar{\mathrm{E}}[v_i^2])(\alpha^*-\alpha_0) + \mathbb{E}_n[\psi_{\alpha_0,\beta_0,\theta_0}(y_i,d_i,x_i)] + O_P(\delta_n^{1/2}n^{-1/2}) = O_P(\delta_n^{1/2}n^{-1/2}).$$

Therefore, $|\mathbb{E}_n[\psi_{\check\alpha,\widehat\beta,\widehat\theta}(y_i,d_i,x_i)]| \leqslant |\mathbb{E}_n[\psi_{\alpha^*,\widehat\beta,\widehat\theta}(y_i,d_i,x_i)]| = O_P(\delta_n^{1/2}n^{-1/2})$, so that

$$(f_\epsilon\bar{\mathrm{E}}[v_i^2])(\check\alpha-\alpha_0) = \mathbb{E}_n[\psi_{\alpha_0,\beta_0,\theta_0}(y_i,d_i,x_i)] + O_P(\delta_n^{1/2}n^{-1/2}),$$

which immediately implies that $\sigma_n^{-1}\sqrt{n}(\check\alpha-\alpha_0) \rightsquigarrow N(0,1)$ since by the Lyapunov CLT,

$$(\bar{\mathrm{E}}[v_i^2]/4)^{-1/2}\sqrt{n}\mathbb{E}_n[\psi_{\alpha_0,\beta_0,\theta_0}(y_i,d_i,x_i)] \rightsquigarrow N(0,1).$$

Part 2. (Proof for the second assertion). First consider the denominator of $L_n(\alpha_0)$. We have that

$$\begin{aligned}|\mathbb{E}_n[\widehat v_i^2] - \mathbb{E}_n[v_i^2]| &= |\mathbb{E}_n[(\widehat v_i - v_i)(\widehat v_i + v_i)]| \leqslant \|\widehat v_i - v_i\|_{2,n}\|\widehat v_i + v_i\|_{2,n} \\ &\leqslant \|x_i'(\widehat\theta-\theta_0)\|_{2,n}(2\|v_i\|_{2,n} + \|x_i'(\widehat\theta-\theta_0)\|_{2,n}) \lesssim_P \delta_n,\end{aligned}$$

where we have used the fact that $\|v_i\|_{2,n} \lesssim_P (\bar{\mathrm{E}}[v_i^2])^{1/2} = O(1)$ (which is guaranteed by ILAD(i)).

Next consider the numerator of $L_n(\alpha_0)$. Since $\bar{\mathrm{E}}[\psi_{\alpha_0,\beta_0,\theta_0}(y_i,d_i,x_i)] = 0$ we have

$$\mathbb{E}_n[\psi_{\alpha_0,\widehat\beta,\widehat\theta}(y_i,d_i,x_i)] = n^{-1/2}\mathbb{G}_n(\psi_{\alpha_0,\widehat\beta,\widehat\theta} - \psi_{\alpha_0,\beta_0,\theta_0}) + \Gamma(\alpha_0,\widehat\beta,\widehat\theta) + \mathbb{E}_n[\psi_{\alpha_0,\beta_0,\theta_0}(y_i,d_i,x_i)].$$

By Condition ILAD(iii) and the previous calculation, we have

$$|\mathbb{G}_n(\psi_{\alpha_0,\widehat{\beta},\widehat{\theta}} - \psi_{\alpha_0,\beta_0,\theta_0})| \lesssim_P \delta_n \text{ and } |\Gamma(\alpha_0,\widehat{\beta},\widehat{\theta})| \lesssim_P \delta_n n^{-1/2}.$$

Therefore, using the simple identity that $nA_n^2 = nB_n^2 + n(A_n - B_n)^2 + 2nB_n(A_n - B_n)$ with

$$A_n = \mathbb{E}_n[\psi_{\alpha_0,\widehat{\beta},\widehat{\theta}}(y_i,d_i,x_i)] \text{ and } B_n = \mathbb{E}_n[\psi_{\alpha_0,\beta_0,\theta_0}(y_i,d_i,x_i)] \lesssim_P (\bar{\mathrm{E}}[v_i^2])n^{-1/2},$$

we have

$$nL_n(\alpha_0) = \frac{4n|\mathbb{E}_n[\psi_{\alpha_0,\widehat{\beta},\widehat{\theta}}(y_i,d_i,x_i)]|^2}{\mathbb{E}_n[\widehat{v}_i^2]} = \frac{4n|\mathbb{E}_n[\psi_{\alpha_0,\beta_0,\theta_0}(y_i,d_i,x_i)]|^2}{\bar{\mathrm{E}}[v_i^2]} + O_P(\delta_n)$$

since $\bar{\mathrm{E}}[v_i^2] \geqslant c$ is bounded away from zero. The result then follows since

$$(\bar{\mathrm{E}}[v_i^2]/4)^{-1/2}\sqrt{n}\mathbb{E}_n[\psi_{\alpha_0,\beta_0,\theta_0}(y_i,d_i,x_i)] \rightsquigarrow N(0,1).$$

□

Comment A.2 (On 1-step procedure)**.** An inspection of the proof leads to the following stochastic expansion:

$$\begin{aligned}\mathbb{E}_n[\psi_{\widehat{\alpha},\widehat{\beta},\widehat{\theta}}(y_i,d_i,x_i)] &= -(f_\epsilon\bar{\mathrm{E}}[v_i^2])(\widehat{\alpha}-\alpha_0) + \mathbb{E}_n[\psi_{\alpha_0,\beta_0,\theta_0}(y_i,d_i,x_i)] \\ &\quad + O_P(\delta_n^{1/2}n^{-1/2} + \delta_n n^{-1/4}|\widehat{\alpha}-\alpha_0| + |\widehat{\alpha}-\alpha_0|^2),\end{aligned}$$

where $\widehat{\alpha}$ is any consistent estimator of α_0. Hence provided that $|\widehat{\alpha}-\alpha_0| = o_P(n^{-1/4})$, the remainder term in the above expansion is $o_P(n^{-1/2})$, and the 1-step estimator $\check{\alpha}$ defined by

$$\check{\alpha} = \widehat{\alpha} + (\mathbb{E}_n[f_\epsilon \widehat{v}_i^2])^{-1}\mathbb{E}_n[\psi_{\widehat{\alpha},\widehat{\beta},\widehat{\theta}}(y_i,d_i,x_i)]$$

has the following stochastic expansion:

$$\begin{aligned}\check{\alpha} &= \widehat{\alpha} + \{f_\epsilon\bar{\mathrm{E}}[v_i^2] + o_P(n^{-1/4})\}^{-1}\{-(f_\epsilon\bar{\mathrm{E}}[v_i^2])(\widehat{\alpha}-\alpha_0) + \mathbb{E}_n[\psi_{\alpha_0,\beta_0,\theta_0}(y_i,d_i,x_i)] + o_P(n^{-1/2})\} \\ &= \alpha_0 + (f_\epsilon\bar{\mathrm{E}}[v_i^2])^{-1}\mathbb{E}_n[\psi_{\alpha_0,\beta_0,\theta_0}(y_i,d_i,x_i)] + o_P(n^{-1/2}),\end{aligned}$$

so that $\sigma_n^{-1}\sqrt{n}(\check{\alpha}-\alpha_0) \rightsquigarrow N(0,1)$.

Appendix B. Proof of Theorem 1

The proof of Theorem 1 uses the properties of Post-ℓ_1-LAD and Post-Lasso. We will collect these properties together with required regularity conditions in Appendix D.

Proof of Theorem 1. We will verify Condition ILAD and the desired result then follows from Lemma 1. The assumptions on the error density $f_\epsilon(\cdot)$ in Condition ILAD(i) are assumed in Condition I(iv). The moment conditions on d_i and v_i in Condition ILAD(i) are assumed in Condition I(ii).

Condition SE implies that $\kappa_{\mathbf{c}}$ is bounded away from zero with probability $1-\Delta_n$ for n sufficiently large, see [9]. Step 1 relies on Post-ℓ_1-LAD. By assumption with probability $1-\Delta_n$ we have $\widehat{s} = \|\widetilde{\beta}\|_0 \leqslant Cs$. Thus, by Condition SE $\phi_{\min}(\widehat{s}+s)$ is bounded away from zero since $\widehat{s}+s \leqslant \ell_n s$ for large enough n with probability $1-\Delta_n$. Moreover, Condition PLAD in Appendix D is implied by Condition I. The required side condition of Lemma 4 is satisfied by relations (F.41) and (F.42). By Lemma 4 we have $|\widehat{\alpha}-\alpha_0| \lesssim_P \sqrt{s\log(p\vee n)/n} \leqslant o(1)\log^{-1} n$ under $s^3\log^3(p\vee n) \leqslant \delta_n n$. Note that this implies

$\{\alpha : |\alpha - \alpha_0| \leqslant n^{-1/2}\log n\} \subset \mathcal{A}$ (with probability $1 - o(1)$) which is required in ILAD(ii) and the (shrinking) definition of $\mathcal{A}$ establishes the initial rate of ILAD(iv). By Lemma 5 in Appendix D we have $\|x_i'(\widetilde{\beta} - \beta_0)\|_{2,n} \lesssim_P \sqrt{s\log(n \vee p)/n}$ since the required side condition holds. Indeed, for $\tilde{x}_i = (d_i, x_i')'$ and $\delta = (\delta_d, \delta_x')'$, because of Condition SE and the fact that $\mathbb{E}_n[|d_i|^3] \lesssim_P \bar{\mathrm{E}}[|d_i|^3] = O(1)$,

$$\begin{aligned}
\inf_{\|\delta\|_0 \leqslant s+Cs} \frac{\|\tilde{x}_i'\delta\|_{2,n}^3}{\mathbb{E}_n[|\tilde{x}_i'\delta|^3]} &\geqslant \inf_{\|\delta\|_0 \leqslant s+Cs} \frac{\{\phi_{\min}(s+Cs)\}^{3/2}\|\delta\|^3}{4\mathbb{E}_n[|x_i'\delta_x|^3] + 4|\delta_d|^3\mathbb{E}_n[|d_i|^3]} \\
&\geqslant \inf_{\|\delta\|_0 \leqslant s+Cs} \frac{\{\phi_{\min}(s+Cs)\}^{3/2}\|\delta\|^3}{4K_x\|\delta_x\|_1\phi_{\max}(s+Cs)\|\delta_x\|^2 + 4\|\delta\|^3\mathbb{E}_n[|d_i|^3]} \\
&\geqslant \frac{\{\phi_{\min}(s+Cs)\}^{3/2}}{4K_x\sqrt{s+Cs}\phi_{\max}(s+Cs) + 4\mathbb{E}_n[|d_i|^3]} \gtrsim_P \frac{1}{K_x\sqrt{s}}.
\end{aligned}$$

Therefore, since $K_x^2 s^2 \log^2(p \vee n) \leqslant \delta_n n$ and $\lambda \lesssim \sqrt{n\log(p \vee n)}$ we have

$$\frac{n\sqrt{\phi_{\min}(s+Cs)}}{\lambda\sqrt{s} + \sqrt{sn\log(p \vee n)}} \inf_{\|\delta\|_0 \leqslant s+Cs} \frac{\|\tilde{x}_i'\delta\|_{2,n}^3}{\mathbb{E}_n[|\tilde{x}_i'\delta|^3]} \gtrsim_P \frac{\sqrt{n}}{K_x s\log(p \vee n)} \to \infty.$$

Step 2 relies on Post-Lasso. Condition HL in Appendix D is implied by Condition I and Lemma 2 applied twice with $\zeta_i = v_i$ and $\zeta_i = d_i$ under the condition that $K_x^4 \log p \leqslant \delta_n n$. By Lemma 7 in Appendix D we have $\|x_i'(\widetilde{\theta} - \theta_0)\|_{2,n} \lesssim_P \sqrt{s\log(n \vee p)/n}$ and $\|\widetilde{\theta}\|_0 \lesssim s$ with probability $1 - o(1)$.

The rates established above for $\widetilde{\theta}$ and $\widetilde{\beta}$ imply (A.25) in ILAD(iii) since by Condition I(ii) $\mathrm{E}[|v_i|] \leqslant (\mathrm{E}[v_i^2])^{1/2} = O(1)$ and $\max_{1 \leqslant i \leqslant n} |x_i'(\widetilde{\theta} - \theta_0)| \lesssim_P K_x\sqrt{s^2\log(p \vee n)/n} = o(1)$.

We now verify the last requirement in Condition ILAD(iii). Consider the following class of functions

$$\mathcal{F}_s = \{(y, d, x) \mapsto 1\{y \leqslant x'\beta + d\alpha\} : \alpha \in \mathbb{R},\ \|\beta\|_0 \leqslant Cs\},$$

which is the union of $\binom{p}{Cs}$ VC-subgraph classes of functions with VC indices bounded by $C's$. Hence

$$\log N(\varepsilon, \mathcal{F}_s, \|\cdot\|_{\mathbb{P}_n,2}) \lesssim s\log p + s\log(1/\varepsilon).$$

Likewise, consider the following class of functions $\mathcal{G}_{s,r} = \{(y, d, x) \mapsto x'\theta : \|\theta\|_0 \leqslant Cs, \|x_i'\theta\|_{2,n} \leqslant r\}$. Then

$$\log N(\varepsilon\|G_{s,r}\|_{\mathbb{P}_n,2}, \mathcal{G}_s, \|\cdot\|_{\mathbb{P}_n,2}) \lesssim s\log p + s\log(1/\varepsilon),$$

where $G_{s,r}(y, d, x) = \max_{\|\theta\|_0 \leqslant Cs, \|x_i'\theta\|_{2,n} \leqslant r} |x'\theta|$.

Note that

$$\begin{aligned}
\sup_{\alpha\in\mathcal{A}} |\mathbb{G}_n(\psi_{\alpha,\widetilde{\beta},\widetilde{\theta}} - \psi_{\alpha,\beta_0,\theta_0})| &\leqslant \sup_{\alpha\in\mathcal{A}} |\mathbb{G}_n(\psi_{\alpha,\widetilde{\beta},\widetilde{\theta}} - \psi_{\alpha,\widetilde{\beta},\theta_0})| && \text{(B.28)} \\
&\quad + \sup_{\alpha\in\mathcal{A}} |\mathbb{G}_n(\psi_{\alpha,\widetilde{\beta},\theta_0} - \psi_{\alpha,\beta_0,\theta_0})|. && \text{(B.29)}
\end{aligned}$$

Consider to bound (B.28). Observe that

$$\psi_{\alpha,\beta,\theta}(y_i, d_i, x_i) - \psi_{\alpha,\beta,\theta_0}(y_i, d_i, x_i) = -(1/2 - 1\{y_i \leqslant x_i'\beta + d_i\alpha\})x_i'(\theta - \theta_0),$$

and consider the class of functions $\mathcal{H}_{s,r}^1 = \{(y, d, x) \mapsto (1/2 - 1\{y \leqslant x'\beta + d\alpha\})x'(\theta - \theta_0) : \alpha \in \mathbb{R},\ \|\beta\|_0 \leqslant Cs,\ \|\theta\|_0 \leqslant Cs,\ \|x_i'(\theta - \theta_0)\|_{2,n} \leqslant r\}$ with $r \lesssim \sqrt{s\log(p \vee n)/n}$. Then by Lemma 9 together with the above entropy calculations (and some straightforward algebras), we have

$$\sup_{g \in \mathcal{H}_{s,r}^1} |\mathbb{G}_n(g)| \lesssim_P \sqrt{s\log(p \vee n)}\sqrt{s\log(p \vee n)/n} = o_P(1),$$

where $s^2\log^2(p\vee n)\leqslant \delta_n n$ is used. Since $\|x_i'(\widetilde\theta-\theta_0)\|_{2,n}\lesssim_P \sqrt{s\log(n\vee p)/n}$ and $\|\widetilde\beta\|_0\vee\|\widetilde\theta\|_0\lesssim s$ with probability $1-o(1)$, we conclude that (B.28) $=o_P(1)$.

Lastly consider to bound (B.29). Observe that

$$\psi_{\alpha,\beta,\theta_0}(y_i,d_i,x_i)-\psi_{\alpha,\beta,\theta_0}(y_i,d_i,x_i)=-(1\{y_i\leqslant x_i'\beta+d_i\alpha\}-1\{y_i\leqslant x_i'\beta_0+d_i\alpha\})v_i,$$

where $v_i=d_i-x_i'\theta_0$, and consider the class of functions $\mathcal{H}^2_{s,r}=\{(y,d,x)\mapsto(1\{y\leqslant x'\beta+d\alpha\}-1\{y\leqslant x'\beta_0+d\alpha\})(d-x'\theta_0):\alpha\in\mathbb{R},\ \|\beta\|_0\leqslant Cs,\|x_i'(\beta-\beta_0)\|_{2,n}\leqslant r\}$ with $r\lesssim\sqrt{s\log(p\vee n)/n}$. Then by Lemma 9 together with the above entropy calculations (and some straightforward algebras), we have

$$\sup_{g\in\mathcal{H}^2_{s,r}}|\mathbb{G}_n(g)|\lesssim_P\sqrt{s\log(p\vee n)}\sup_{g\in\mathcal{H}^2_{s,r}}\sqrt{\mathbb{E}_n[g(y_i,d_i,x_i)^2]\vee\bar{\mathrm{E}}[g(y_i,d_i,x_i)^2]}.$$

Here we have

$$\bar{\mathrm{E}}[g(y_i,d_i,x_i)^2]\leqslant C\|x_i'(\beta-\beta_0)\|_{2,n}(\bar{\mathrm{E}}[v_i^4])^{1/2}\lesssim\sqrt{s\log(p\vee n)/n}.$$

On the other hand,

$$\sup_{g\in\mathcal{H}^2_{s,r}}\mathbb{E}_n[g(y_i,d_i,x_i)^2]\leqslant n^{-1/2}\sup_{g\in\mathcal{H}^2_{s,r}}\mathbb{G}_n(g^2)+\sup_{g\in\mathcal{H}^2_{s,r}}\bar{\mathrm{E}}[g(y_i,d_i,x_i)^2],\tag{B.30}$$

and apply Lemma 9 to the first term on the right side of (B.30). Then we have

$$\begin{aligned}\sup_{g\in\mathcal{H}^2_{s,r}}\mathbb{G}_n(g^2)&\lesssim_P\sqrt{s\log(p\vee n)}\sup_{g\in\mathcal{H}^2_{s,r}}\sqrt{\mathbb{E}_n[g(y_i,d_i,x_i)^4]\vee\bar{\mathrm{E}}[g(y_i,d_i,x_i)^4]}\\&\lesssim\sqrt{s\log(p\vee n)}\sqrt{\mathbb{E}_n[v_i^4]\vee\bar{\mathrm{E}}[v_i^4]}\lesssim_P\sqrt{s\log(p\vee n)}\sqrt{\bar{\mathrm{E}}[v_i^4]}.\end{aligned}$$

Since $\|x_i'(\widetilde\beta-\beta_0)\|_{2,n}\lesssim_P\sqrt{s\log(n\vee p)/n}$ and $\|\widetilde\beta\|_0\leqslant Cs$ with probability $1-\Delta_n$, we conclude that

$$\text{(B.29)}\lesssim_P\sqrt{s\log(p\vee n)}(s\log(p\vee n)/n)^{1/4}=o(1),$$

where $s^3\log^3(p\vee n)\leqslant\delta_n n$ is used. □

Appendix C. Auxiliary Technical Results

In this section we collect two auxiliary technical results. Their proofs are given in the supplementary appendix.

Lemma 2. *Let $x_1,\dots,x_n$ be non-stochastic vectors in $\mathbb{R}^p$ with $\max_{1\leqslant i\leqslant n}\|x_i\|_\infty\leqslant K_x$. Let $\zeta_1,\dots,\zeta_n$ be independent random variables such that $\bar{\mathrm{E}}[|\zeta_i|^q]<\infty$ for some $q\geqslant 4$. Then with probability at least $1-8\tau$,*

$$\max_{1\leqslant j\leqslant p}|(\mathbb{E}_n-\bar{\mathrm{E}})[x_{ij}^2\zeta_i^2]|\leqslant 4\sqrt{\frac{\log(2p/\tau)}{n}}K_x^2(\bar{\mathrm{E}}[|\zeta_i|^q]/\tau)^{4/q}.$$

Lemma 3. *Let $T=\operatorname{support}(\beta_0)$, $|T|=\|\beta_0\|_0\leqslant s$ and $\|\widehat\beta_{T^c}\|_1\leqslant\mathbf{c}\|\widehat\beta_T-\beta_0\|_1$. Moreover, let $\widehat\beta^{(2m)}$ denote the vector formed by the largest $2m$ components of $\widehat\beta$ in absolute value and zero in the remaining components. Then for $m\geqslant s$ we have that $\widehat\beta^{(2m)}$ satisfies*

$$\|x_i'(\widehat\beta^{(2m)}-\beta_0)\|_{2,n}\leqslant\|x_i'(\widehat\beta-\beta_0)\|_{2,n}+\sqrt{\phi_{\max}(m)/m}\ \mathbf{c}\|\widehat\beta_T-\beta_0\|_1,$$

where $\phi_{\max}(m)/m\leqslant 2\phi_{\max}(s)/s$ and $\|\widehat\beta_T-\beta_0\|_1\leqslant\sqrt{s}\|x_i'(\widehat\beta-\beta_0)\|_{2,n}/\kappa_{\mathbf{c}}$.

References

[1] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28:253–263, 2008.

[2] A. Belloni, D. Chen, V. Chernozhukov, and C. Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2430, November 2012.

[3] A. Belloni and V. Chernozhukov. ℓ_1-penalized quantile regression for high dimensional sparse models. *Annals of Statistics*, 39(1):82–130, 2011.

[4] A. Belloni and V. Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013.

[5] A. Belloni, V. Chernozhukov, and I. Fernandez-Val. Conditional Quantile Processes based on Series or Many Regressors. *ArXiv e-prints*, May 2011.

[6] A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection amongst high-dimensional controls. *ArXiv*, 2011.

[7] A. Belloni, V. Chernozhukov, and C. Hansen. Inference for high-dimensional sparse econometric models. *Advances in Economics and Econometrics. 10th World Congress of Econometric Society, held in August 2010*, III:245–295, 2013.

[8] A. Belloni, V. Chernozhukov, and K. Kato. Robust inference in high-dimensional sparse quantile regression models. *Working Paper*, 2013.

[9] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

[10] Victor Chernozhukov and Christian Hansen. Instrumental variable quantile regression: A robust inference approach. *Journal of Econometrics*, 142:379–398, 2008.

[11] Xuming He and Qi-Man Shao. On parameters of increasing dimensions. *J. Multivariate Anal.*, 73(1):120–135, 2000.

[12] Bing-Yi Jing, Qi-Man Shao, and Qiying Wang. Self-normalized Cramer-type large deviations for independent random variables. *Ann. Probab.*, 31(4):2167–2215, 2003.

[13] K. Kato. Group lasso for high dimensional sparse quantile regression models. Preprint, ArXiv, 2011.

[14] Roger Koenker. *Quantile regression*, volume 38 of *Econometric Society Monographs*. Cambridge University Press, Cambridge, 2005.

[15] Michael R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Series in Statistics. Springer, Berlin, 2008.

[16] Sokbae Lee. Efficient semiparametric estimation of a partially linear quantile regression model. *Econometric theory*, 19:1–31, 2003.

[17] Hannes Leeb and Benedikt M. Pötscher. Model selection and inference: facts and fiction. *Economic Theory*, 21:21–59, 2005.

[18] Hannes Leeb and Benedikt M. Pötscher. Sparse estimators and the oracle property, or the return of Hodges' estimator. *J. Econometrics*, 142(1):201–211, 2008.

[19] J. Neyman. Optimal asymptotic tests of composite statistical hypotheses. In U. Grenander, editor, *Probability and Statistics, the Harold Cramer Volume*. New York: John Wiley and Sons, Inc., 1959.

[20] J. Neyman. $c(\alpha)$ tests and their use. *Sankhya*, 41:1–21, 1979.

[21] J. L. Powell. Censored regression quantiles. *Journal of Econometrics*, 32:143–155, 1986.

[22] M. Rudelson and S. Zhou. Reconstruction from anisotropic random measurements. *ArXiv:1106.1151*, 2011.

[23] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58:267–288, 1996.

[24] S. A. van de Geer. High-dimensional generalized linear models and the lasso. *Annals of Statistics*, 36(2):614–645, 2008.

[25] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics, 1996.

[26] Lie Wang. L_1 penalized lad estimator for high dimensional ilnear regression. *ArXiv*, 2012.

[27] Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low-dimensional parameters with high-dimensional data. *ArXiv.org*, (arXiv:1110.2563v1), 2011.

[28] S. Zhou. Restricted eigenvalue conditions on subgaussian matrices. *ArXiv:0904.4723v2*, 2009.

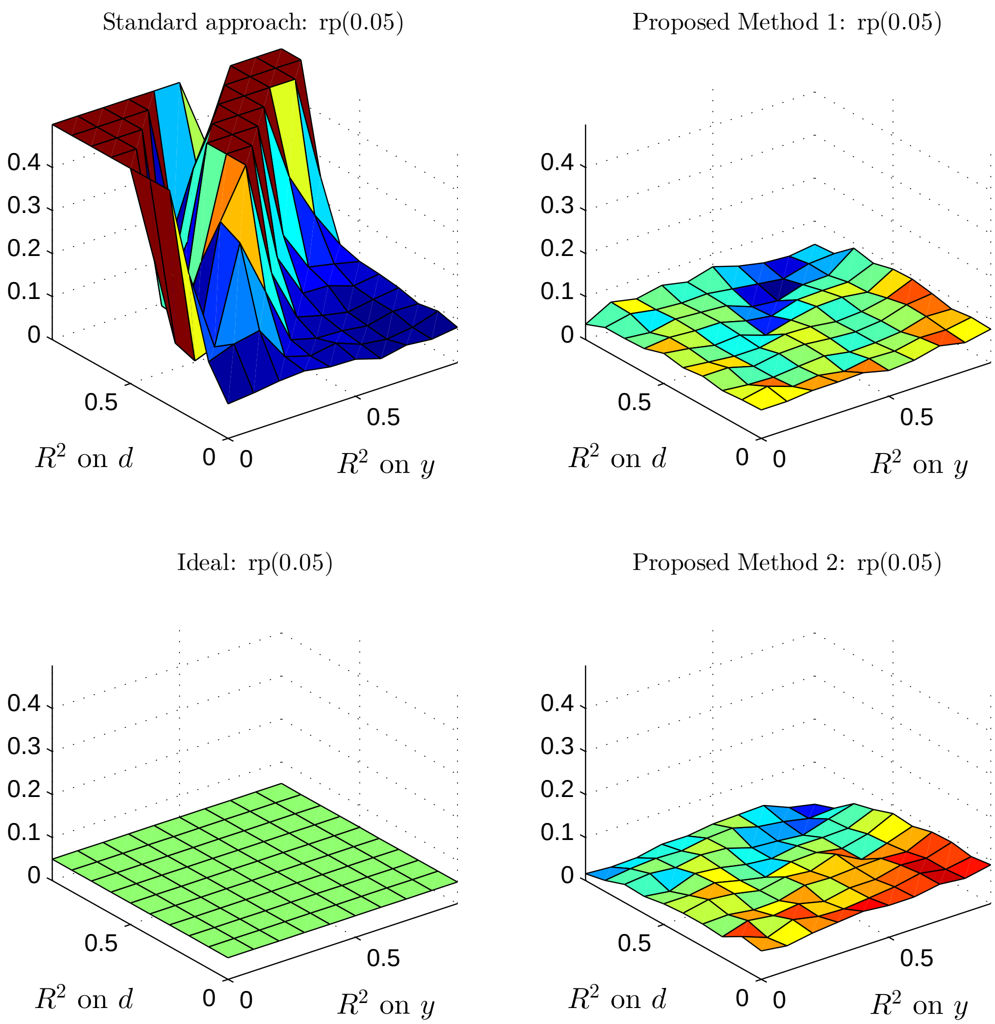

FIGURE 1. The figure displays the empirical rejection probabilities of the nominal 5% level tests of a true hypothesis based on different testing procedures: the top left plot is based on the standard post-model selection procedure based on $\widetilde{\alpha}$, the top right plot is based on the proposed post-model selection procedure based on $\check{\alpha}$, and the bottom left plot is based on another proposed procedure based on the statistic L_n. The results are based on 500 replications for each of the 100 combinations of R^2's in the primary and auxiliary equations in (3.14). Ideally we should observe the 5% rejection rate (of a true null) uniformly across the parameter space (as in bottom right plot).

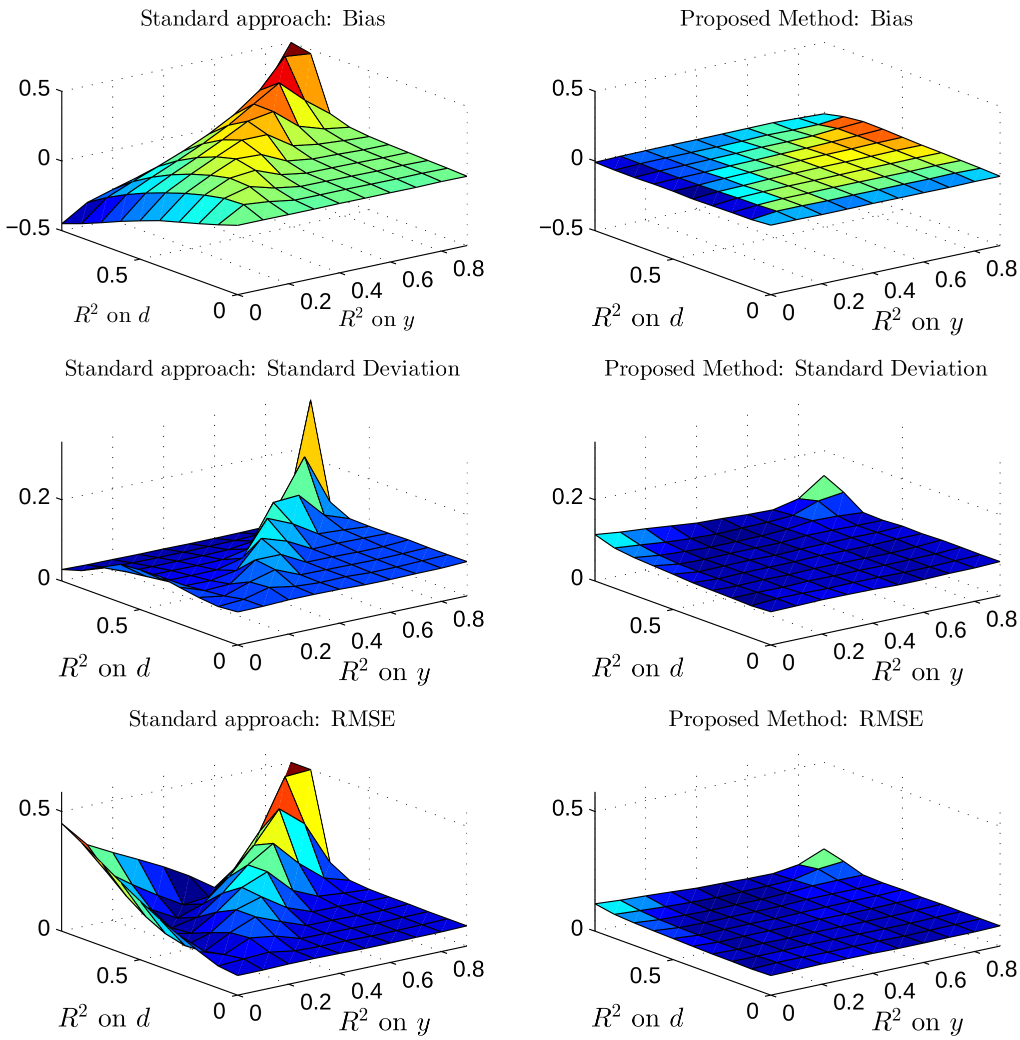

FIGURE 2. The figure displays mean bias (top row), standard deviation (middle row), and root mean square error (bottom row) for the the proposed post-model selection estimator $\check{\alpha}$ (right column) and the standard post-model selection estimator $\widetilde{\alpha}$ (left column). The results are based on 500 replications for each of the 100 combinations of R^2's in the primary and auxiliary equations in (3.14).

Supplementary Appendix for "Uniform Post Selection Inference for LAD Regression Models"

Appendix D. Auxiliary Results for ℓ_1-LAD and Heteroscedastic Lasso

In this section we state relevant theoretical results on the performance of the estimators ℓ_1-LAD, Post-ℓ_1-LAD, heteroscedastic Lasso, and heteroscedastic Post-Lasso. There results were developed in [3] and [2]. The main design condition relies on the restricted eigenvalue proposed in [9], namely for $\tilde{x}_i = (d_i, x_i')'$

$$\kappa_{\mathbf{c}} = \inf_{\|\delta_{T^c}\|_1 \leqslant \mathbf{c}\|\delta_T\|_1} \|\tilde{x}_i'\delta\|_{2,n}/\|\delta_T\|, \tag{D.31}$$

where $\mathbf{c} = (c+1)/(c-1)$ for the slack constant $c > 1$, see [9]. It is well known that Condition SE implies that $\kappa_{\mathbf{c}}$ is bounded away from zero if $\mathbf{c}$ is bounded for any subset $T \subset \{1, \ldots, p\}$ with $|T| \leqslant s$.

D.1. ℓ_1-Penalized LAD.

For a data generating process such that $\mathrm{P}(y_i \leqslant \tilde{x}_i'\eta_0 \mid \tilde{x}_i) = 1/2$, independent across i ($i = 1, \ldots, n$) we consider the estimation of η_0 via the ℓ_1-penalized LAD regression estimate

$$\widehat{\eta} \in \arg\min_{\eta} \mathbb{E}_n[|y_i - \tilde{x}_i'\eta|] + \frac{\lambda}{n}\|\eta\|_1.$$

As established in [3] and [26], under the event that

$$\frac{\lambda}{n} \geqslant 2c\|\mathbb{E}_n[(1/2 - 1\{y_i \leqslant \tilde{x}_i'\eta_0\})\tilde{x}_i]\|_\infty, \tag{D.32}$$

the estimator above achieves good theoretical guarantees under mild design conditions. Although η_0 is unknown, we can set λ so that the event in (D.32) holds with high probability. In particular, the pivotal rule discussed in [3] proposes to set $\lambda = c'n\Lambda(1-\gamma \mid \tilde{x})$ for $c' > c$ and $\gamma \to 0$ where

$$\Lambda(1-\gamma \mid \tilde{x}) := (1-\gamma)\text{-quantile of } 2\|\mathbb{E}_n[(1/2 - 1\{U_i \leqslant 1/2\})\tilde{x}_i]\|_\infty, \tag{D.33}$$

and where U_i are independent uniform random variables on $(0,1)$, independent of $\tilde{x}_1, \ldots, \tilde{x}_n$. We suggest $\gamma = 0.1/\log n$ and $c' = 1.1c$. This quantity can be easily approximated via simulations. Below we summarize required regularity conditions.

Condition PLAD. Assume that $\|\eta_0\|_0 = s \geqslant 1$, $\mathbb{E}_n[\tilde{x}_{ij}^2] = 1$ for all $1 \leqslant j \leqslant p$, the conditional density of y_i given d_i, denoted by $f_i(\cdot)$, and its derivative are bounded by $\bar{f}$ and $\bar{f}'$, respectively, and $f_i(\tilde{x}_i'\eta_0) \geqslant \underline{f} > 0$ is bounded away from zero uniformly in n.

Condition PLAD is implied by Condition I. The assumption on the conditional density is standard in the quantile regression literature even with fixed p or p increasing slower than n (see respectively [14] and [5]). Next we present bounds on the prediction norm of the ℓ_1-LAD estimator.

Lemma 4 (Estimation Error of ℓ_1-LAD)**.** *Under Condition PLAD, and using $\lambda = c'n\Lambda(1-\gamma \mid \tilde{x})$, we have with probability $1 - 2\gamma - o(1)$ for n large enough*

$$\|\tilde{x}_i'(\widehat{\eta} - \eta_0)\|_{2,n} \lesssim \frac{\lambda\sqrt{s}}{n\kappa_{\mathbf{c}}} + \frac{1}{\kappa_{\mathbf{c}}}\sqrt{\frac{s\log(p/\gamma)}{n}},$$

provided $\left\{ \frac{n\kappa_{\mathbf{c}}}{\lambda\sqrt{s}} + \frac{n\kappa_{\mathbf{c}}}{\sqrt{sn\log([p\vee n]/\gamma)}} \right\} \frac{\bar{f}\bar{f}'}{\underline{f}} \inf_{\delta\in\Delta_{\mathbf{c}}} \frac{\|\tilde{x}_i'\delta\|_{2,n}^3}{\mathbb{E}_n[|\tilde{x}_i'\delta|^3]} \to \infty.$

Lemma 4 establishes the rate of convergence in the prediction norm for the ℓ_1-LAD estimator in a parametric setting. The extra growth condition required for identification is mild. For instance we typically have $\lambda \lesssim \sqrt{\log(n\vee p)/n}$ and for many designs of interest we have $\inf_{\delta\in\Delta_{\mathbf{c}}} \|\tilde{x}_i'\delta\|_{2,n}^3/\mathbb{E}_n[|\tilde{x}_i'\delta|^3]$ bounded away from zero (see [3]). For more general designs we have

$$\inf_{\delta\in\Delta_{\mathbf{c}}} \frac{\|\tilde{x}_i'\delta\|_{2,n}^3}{\mathbb{E}_n[|\tilde{x}_i'\delta|^3]} \geqslant \inf_{\delta\in\Delta_{\mathbf{c}}} \frac{\|\tilde{x}_i'\delta\|_{2,n}}{\|\delta\|_1 \max_{i\leqslant n}\|\tilde{x}_i\|_\infty} \geqslant \frac{\kappa_{\mathbf{c}}}{\sqrt{s}(1+\mathbf{c})\max_{i\leqslant n}\|\tilde{x}_i\|_\infty}$$

which implies the extra growth condition under $K_x^2 s^2\log(p\vee n) \leqslant \delta_n\kappa_{\mathbf{c}}^2 n$.

In order to alleviate the bias introduced by the ℓ_1-penalty, we can consider the associated post-model selection estimate associated with a selected support $\widehat{T}$

$$\widetilde{\eta} \in \arg\min_{\eta} \left\{ \mathbb{E}_n[|y_i - \tilde{x}_i'\eta|] : \eta_j = 0 \text{ if } j \notin \widehat{T} \right\}. \tag{D.34}$$

The following result characterizes the performance of the estimator in (D.34), see [3] for the proof.

Lemma 5 (Estimation Error of Post-ℓ_1-LAD)**.** *Assume the conditions of Lemma 4 hold,* support$(\widehat{\eta}) \subseteq \widehat{T}$*, and let* $\widehat{s} = |\widehat{T}|$*. Then we have for* n *large enough*

$$\|\tilde{x}_i'(\widetilde{\eta}-\eta_0)\|_{2,n} \lesssim_P \sqrt{\frac{(\widehat{s}+s)\log(n\vee p)}{n\phi_{\min}(\widehat{s}+s)}} + \frac{\lambda\sqrt{s}}{n\kappa_{\mathbf{c}}} + \frac{1}{\kappa_{\mathbf{c}}}\sqrt{\frac{s\log(p/\gamma)}{n}},$$

provided $\left\{ \frac{n\sqrt{\phi_{\min}(\widehat{s}+s)}}{\lambda\sqrt{s}} + \frac{n\sqrt{\phi_{\min}(\widehat{s}+s)}}{\sqrt{sn\log([p\vee n]/\gamma)}} \right\} \frac{\bar{f}\bar{f}'}{\underline{f}} \inf_{\|\delta\|_0\leqslant\widehat{s}+s} \frac{\|\tilde{x}_i'\delta\|_{2,n}^3}{\mathbb{E}_n[|\tilde{x}_i'\delta|^3]} \to_P \infty.$

Lemma 5 provides the rate of convergence in the prediction norm for the post model selection estimator despite of possible imperfect model selection. The rates rely on the overall quality of the selected model (which is at least as good as the model selected by ℓ_1-LAD) and the overall number of components $\widehat{s}$. Once again the extra growth condition required for identification is mild. For more general designs we have

$$\inf_{\|\delta\|_0\leqslant\widehat{s}+s} \frac{\|\tilde{x}_i'\delta\|_{2,n}^3}{\mathbb{E}_n[|\tilde{x}_i'\delta|^3]} \geqslant \inf_{\|\delta\|_0\leqslant\widehat{s}+s} \frac{\|\tilde{x}_i'\delta\|_{2,n}}{\|\delta\|_1\max_{i\leqslant n}\|\tilde{x}_i\|_\infty} \geqslant \frac{\sqrt{\phi_{\min}(\widehat{s}+s)}}{\sqrt{\widehat{s}+s}\max_{i\leqslant n}\|\tilde{x}_i\|_\infty}.$$

Comment D.1. In Step 1 of Algorithm 2 we use ℓ_1-LAD with $\tilde{x}_i = (d_i, x_i')'$, $\widehat{\delta} := \widehat{\eta}-\eta_0 = (\widehat{\alpha}-\alpha_0, \widehat{\beta}'-\beta_0')'$, and we are interested on rates for $\|x_i'(\widehat{\beta}-\beta_0)\|_{2,n}$ instead of $\|\tilde{x}_i'\widehat{\delta}\|_{2,n}$. However, it follows that

$$\|x_i'(\widehat{\beta}-\beta_0)\|_{2,n} \leqslant \|\tilde{x}_i'\widehat{\delta}\|_{2,n} + |\widehat{\alpha}-\alpha_0|\cdot\|d_i\|_{2,n}.$$

Since $s \geqslant 1$, without loss of generality we can assume the component associated with the treatment d_i belongs to T (at the cost of increasing the cardinality of T by one which will not affect the rate of convergence). Therefore we have that

$$|\widehat{\alpha}-\alpha_0| \leqslant \|\widehat{\delta}_T\| \leqslant \|\tilde{x}_i'\widehat{\delta}\|_{2,n}/\kappa_{\mathbf{c}}.$$

In most applications of interest $\|d_i\|_{2,n}$ and $1/\kappa_{\mathbf{c}}$ are bounded from above with high probability. Similarly, in Step 1 of Algorithm 1 we have that the Post-ℓ_1-LAD estimator satisfies

$$\|x_i'(\widetilde{\beta}-\beta_0)\|_{2,n} \leqslant \|\tilde{x}_i'\widetilde{\delta}\|_{2,n}\left(1 + \|d_i\|_{2,n}/\sqrt{\phi_{\min}(\widehat{s}+s)}\right).$$

D.2. **Heteroscedastic Lasso.** In this section we consider the equation (1.4) in the form

$$d_i = x_i'\theta_0 + v_i, \ \mathrm{E}[v_i] = 0, \tag{D.35}$$

where we observe $\{(d_i, x_i')'\}_{i=1}^n$, $(x_i)_{i=1}^n$ are non-stochastic and normalized in such a way that $\mathbb{E}_n[x_{ij}^2] = 1$, for all $1 \leqslant j \leqslant p$, and $(v_i)_{i=1}^n$ are independent across i but not necessary identically distributed. The unknown support of θ_0 is denoted by T_d and it satisfies $|T_d| \leqslant s$. To estimate θ_0 and consequently v_i, we compute

$$\widehat{\theta} \in \arg\min_{\theta} \mathbb{E}_n[(d_i - x_i'\theta)^2] + \frac{\lambda}{n}\|\widehat{\Gamma}\theta\|_1 \text{ and set } \widehat{v}_i = d_i - x_i'\widehat{\theta}, \quad i = 1, \ldots, n, \tag{D.36}$$

where λ and $\widehat{\Gamma}$ are the associated penalty level and loadings which are potentially data-driven. In this case the following regularization event plays an important role

$$\frac{\lambda}{n} \geqslant 2c\|\widehat{\Gamma}^{-1}\mathbb{E}_n[x_i(d_i - x_i'\theta_0)]\|_\infty. \tag{D.37}$$

As discussed in [9], [4] and [2], the event above implies that the estimator $\widehat{\theta}$ satisfies $\|\widehat{\theta}_{T_d^c}\|_1 \leqslant \mathbf{c}\|\widehat{\theta}_{T_d} - \theta_0\|_1$ where $\mathbf{c} = (c+1)/(c-1)$. Thus rates of convergence for $\widehat{\theta}$ and $\widehat{v}_i$ defined on (D.36) can be established based on the restricted eigenvalue $\kappa_{\mathbf{c}}$ defined in (D.31) with $\tilde{x}_i = x_i$ and $T = T_d$.

The following are sufficient high-level conditions where again the sequences Δ_n and δ_n go to zero and C is a positive constant independent of n.

Condition HL. For the model (D.35), suppose that for $s = s_n \geqslant 1$ we have $\|\theta_0\|_0 \leqslant s$ and
(i) $\max_{1 \leqslant j \leqslant p} (\bar{\mathrm{E}}[|x_{ij}v_i|^3])^{1/3}/(\bar{\mathrm{E}}[|x_{ij}v_i|^2])^{1/2} \leqslant C$ and $\Phi^{-1}(1-\gamma/2p) \leqslant \delta_n n^{1/3}$,
(ii) $\max_{1 \leqslant j \leqslant p} |(\mathbb{E}_n - \bar{\mathrm{E}})[x_{ij}^2 v_i^2]| + \max_{j \leqslant p} |(\mathbb{E}_n - \bar{\mathrm{E}})[x_{ij}^2 d_i^2]| \leqslant \delta_n$, with probability $1 - \Delta_n$.

Condition HL is implied by Conditions I and growth conditions (see Lemma 2). Several primitive moment conditions imply the various cross moments bounds. These conditions also allow us to invoke moderate deviation theorems for self-normalized sums from [12] to bound some important error components. Despite heteroscedastic non-Gaussian noise, Those results allows a sharp choice of penalty level and loadings was analyzed in [2] which is summarized by the following lemma.

Valid options for setting the penalty level and the loadings for $j = 1, \ldots, p$, are

$$\begin{array}{llll} \text{initial} & \widehat{\gamma}_j = \sqrt{\mathbb{E}_n[x_{ij}^2(d_i - \bar{d})^2]}, & \lambda = 2c\sqrt{n}\Phi^{-1}(1-\gamma/(2p)), \\ \text{refined} & \widehat{\gamma}_j = \sqrt{\mathbb{E}_n[x_{ij}^2\widehat{v}_i^2]}, & \lambda = 2c\sqrt{n}\Phi^{-1}(1-\gamma/(2p)), \end{array} \tag{D.38}$$

where $c > 1$ is a constant, $\gamma \in (0,1)$, $\bar{d} := \mathbb{E}_n[d_i]$ and $\widehat{v}_i$ is an estimate of v_i based on Lasso with the initial option (or iterations). [2] established that using either of the choices in (D.38) implies that the regularization event (D.37) holds with high probability. Next we present results on the performance of the estimators generated by Lasso.

Lemma 6. *Under Condition HL and setting $\lambda = 2c'\sqrt{n}\Phi^{-1}(1-\gamma/2p)$ for $c' > c > 1$, and using penalty loadings as in (D.38), there is an uniformly bounded $\mathbf{c}$ such that we have*

$$\|\widehat{v}_i - v_i\|_{2,n} = \|x_i'(\widehat{\theta} - \theta_0)\|_{2,n} \lesssim_P \frac{\lambda\sqrt{s}}{n\kappa_{\mathbf{c}}} \qquad \textit{and} \qquad \|\widehat{v}_i - v_i\|_\infty \leqslant \|\widehat{\theta} - \theta_0\|_1 \max_{i \leqslant n} \|x_i\|_\infty.$$

Associated with Lasso we can define the Post-Lasso estimator as

$$\widetilde{\theta} \in \arg\min_{\theta} \left\{ \mathbb{E}_n[(d_i - x_i'\theta)^2] : \theta_j = 0 \text{ if } \widehat{\theta}_j = 0 \right\} \text{ and set } \tilde{v}_i = d_i - x_i'\widetilde{\theta}. \tag{D.39}$$

That is, the Post-Lasso estimator is simply the least squares estimator applied to the covariates selected by Lasso in (D.36). Sparsity properties of the Lasso estimator $\widehat{\theta}$ under estimated weights follows similarly to the standard Lasso analysis derived in [2]. By combining such sparsity properties and the rates in the prediction norm we can establish rates for the post-model selection estimator under estimated weights. The following result summarizes the properties of the Post-Lasso estimator.

Lemma 7 (Model Selection Properties of Lasso and Properties of Post-Lasso)**.** *Suppose that Conditions HL and SE hold. Consider the Lasso estimator with penalty level and loadings specified as in Lemma 6. Then the data-dependent model $\widehat{T}_d$ selected by the Lasso estimator $\widehat{\theta}$ satisfies with probability $1-\Delta_n$:*

$$\|\widetilde{\theta}\|_0 = |\widehat{T}_d| \lesssim s. \tag{D.40}$$

Moreover, the Post-Lasso estimator obeys

$$\|\tilde{v}_i - v_i\|_{2,n} = \|x_i'(\widetilde{\theta} - \theta_0)\|_{2,n} \lesssim_P \sqrt{\frac{s\log(p \vee n)}{n}}.$$

Appendix E. Alternative Implementation via Double Selection

An alternative proposal for the method is reminiscent of the double selection method proposed in [6] for partial linear models. This version replaces Step 3 with a LAD regression of y on d and all covariates selected in Steps 1 and 2 (i.e. the union of the selected sets). The method is described as follows:

Algorithm 3. (A Double Selection Method)

Step 1 Run Post-ℓ_1-LAD of y_i on d_i and x_i:

$$(\widehat{\alpha}, \widehat{\beta}) \in \arg\min_{\alpha,\beta} \mathbb{E}_n[|y_i - d_i\alpha - x_i'\beta|] + \frac{\lambda_1}{n}\|(\alpha,\beta)\|_1.$$

Step 2 Run Heteroscedastic Lasso of d_i on x_i:

$$\widehat{\theta} \in \arg\min_{\theta} \mathbb{E}_n[(d_i - x_i'\theta)^2] + \frac{\lambda_2}{n}\|\widehat{\Gamma}\theta\|_1.$$

Step 3 Run LAD regression of y_i on d_i and the covariates selected in Step 1 and 2:

$$(\check{\alpha}, \check{\beta}) \in \arg\min_{\alpha,\beta} \{\mathbb{E}_n[|y_i - d_i\alpha - x_i'\beta|] : \text{ support}(\beta) \subseteq \text{support}(\widehat{\beta}) \cup \text{support}(\widehat{\theta})\}.$$

The double selection algorithm has three steps: (1) select covariates based on the standard ℓ_1-LAD regression, (2) select covariates based on heteroscedastic Lasso of the treatment equation, and (3) run a LAD regression with the treatment and all selected covariates.

This approach can also be analyzed through Lemma 1 since it creates instruments implicitly. To see that let $\widehat{T}^*$ denote the variables selected in Step 1 and 2: $\widehat{T}^* = \text{support}(\widehat{\beta}) \cup \text{support}(\widehat{\theta})$. By the first order conditions for $(\check{\alpha}, \check{\beta})$ we have

$$\|\mathbb{E}_n[\varphi(y_i - d_i\check{\alpha} - x_i'\check{\beta})(d_i, x_{i\widehat{T}^*}')']\| = O\{(\max_{1\leqslant i\leqslant n}|d_i| + K_x|\widehat{T}^*|^{1/2})(1+|\widehat{T}^*|)/n\},$$

which creates an orthogonal relation to any linear combination of $(d_i, x'_{i\widehat{T}^*})'$. In particular, by taking the linear combination $(d_i, x'_{i\widehat{T}^*})(1, -\widetilde{\theta}'_{\widehat{T}^*})' = d_i - x'_{i\widehat{T}^*}\widetilde{\theta}_{\widehat{T}^*} = d_i - x'_i\widetilde{\theta} = \widehat{z}_i$, which is the instrument in Step 2 of Algorithm 1, we have

$$\mathbb{E}_n[\varphi(y_i - d_i\check{\alpha} - x'_i\check{\beta})\widehat{z}_i] = O\{\|(1, -\widetilde{\theta}')'\|(\max_{1\leqslant i\leqslant n}|d_i| + K_x|\widehat{T}^*|^{1/2})(1 + |\widehat{T}^*|)/n\}.$$

As soon as the right side is $o_P(n^{-1/2})$, the double selection estimator $\check{\alpha}$ approximately minimizes

$$\widetilde{L}_n(\alpha) = \frac{|\mathbb{E}_n[\varphi(y_i - d_i\alpha - x'_i\check{\beta})\widehat{z}_i]|^2}{\mathbb{E}_n[\{\varphi(y_i - d_i\check{\alpha} - x'_i\check{\beta})\}^2\widehat{z}_i^2]},$$

where $\widehat{z}_i$ is the instrument created by Step 2 of Algorithm 1. Thus the double selection estimator can be seen as an iterated version of the method based on instruments where the Step 1 estimate $\widetilde{\beta}$ is updated with $\check{\beta}$.

Appendix F. Proof of Theorem 2

Proof of Theorem 2. We will verify Condition ILAD and the desired then follows from Lemma 1. The assumptions on the error density $f_\epsilon(\cdot)$ in Condition ILAD(i) are assumed in Condition I(iv). The moment conditions on d_i and v_i in Condition ILAD(i) are assumed in Condition I(ii).

Condition SE implies that $\kappa_{\mathbf{c}}$ is bounded away from zero with probability $1 - \Delta_n$ for n sufficiently large, see [9]. Step 1 relies on ℓ_1-LAD. Condition PLAD is implied by Condition I. By Lemma 4 and Comment D.1 we have

$$\|x'_i(\widehat{\beta} - \beta_0)\|_{2,n} \lesssim_P \sqrt{s\log(n \vee p)/n} \ \text{ and } \ |\widehat{\alpha} - \alpha_0| \lesssim_P \sqrt{s\log(p \vee n)/n} \lesssim o(1)\log^{-1} n$$

because $s^3\log^3(n \vee p) \leqslant \delta_n n$ and the required side condition holds. Indeed, without loss of generality assume that T contains the treatment so that for $\tilde{x}_i = (d_i, x'_i)'$, $\delta = (\delta_d, \delta'_x)'$, because of Condition SE and the fact that $\mathbb{E}_n[|d_i|^3] \lesssim_P \bar{\mathrm{E}}[|d_i|^3] = O(1)$, we have

$$\begin{aligned}
\inf_{\delta\in\Delta_{\mathbf{c}}} \tfrac{\|\tilde{x}'_i\delta\|^3_{2,n}}{\mathbb{E}_n[|\tilde{x}'_i\delta|^3]} &\geqslant \inf_{\delta\in\Delta_{\mathbf{c}}} \tfrac{\|\tilde{x}'_i\delta\|^2_{2,n}\|\delta_T\|\kappa_{\mathbf{c}}}{4\mathbb{E}_n[|x'_i\delta_x|^3]+4\mathbb{E}_n[|d_i\delta_d|^3]} \geqslant \inf_{\delta\in\Delta_{\mathbf{c}}} \tfrac{\|\tilde{x}'_i\delta\|^2_{2,n}\|\delta_T\|\kappa_{\mathbf{c}}}{4K_x\|\delta_x\|_1\mathbb{E}_n[|x'_i\delta_x|^2]+4|\delta_d|^3\mathbb{E}_n[|d_i|^3]} \\
&\geqslant \inf_{\delta\in\Delta_{\mathbf{c}}} \tfrac{\|\tilde{x}'_i\delta\|^2_{2,n}\|\delta_T\|\kappa_{\mathbf{c}}}{4K_x\|\delta_x\|_1\{\|\tilde{x}'_i\delta\|_{2,n}+\|\delta_d d_i\|_{2,n}\}^2+4|\delta_d|^2\mathbb{E}_n[|d_i|^3]\|\delta_T\|_1} \\
&\geqslant \inf_{\delta\in\Delta_{\mathbf{c}}} \tfrac{\|\tilde{x}'_i\delta\|^2_{2,n}\|\delta_T\|_1\kappa_{\mathbf{c}}/\sqrt{s}}{8K_x(1+\mathbf{c})\|\delta_T\|_1\|\tilde{x}'_i\delta\|^2_{2,n}+8K_x(1+\mathbf{c})\|\delta_T\|_1|\delta_d|^2\{\|d_i\|^2_{2,n}+\mathbb{E}_n[|d_i|^3]\}} \\
&\geqslant \tfrac{\kappa_{\mathbf{c}}/\sqrt{s}}{8K_x(1+\mathbf{c})\{1+\|d_i\|^2_{2,n}/\kappa^2_{\mathbf{c}}+\mathbb{E}_n[|d_i|^3]/\kappa^2_{\mathbf{c}}\}} \gtrsim_P \tfrac{1}{\sqrt{s}K_x}.
\end{aligned} \tag{F.41}$$

Therefore, since $\lambda \lesssim \sqrt{n\log(p \vee n)}$ we have

$$\frac{n\kappa_{\mathbf{c}}}{\lambda\sqrt{s} + \sqrt{sn\log(p \vee n)}} \inf_{\delta\in\Delta_{\mathbf{c}}} \frac{\|\tilde{x}'_i\delta\|^3_{2,n}}{\mathbb{E}_n[|\tilde{x}'_i\delta|^3]} \gtrsim_P \frac{\sqrt{n}}{K_x s\log(p \vee n)} \to_P \infty \tag{F.42}$$

under $K_x^2 s^2\log^2(p \vee n) \leqslant \delta_n n$. Note that the rate for $\widehat{\alpha}$ and the definition of $\mathcal{A}$ implies $\{\alpha : |\alpha - \alpha_0| \leqslant n^{-1/2}\log n\} \subset \mathcal{A}$ (with probability $1 - o(1)$) which is required in ILAD(ii). Moreover, by the (shrinking) definition of $\mathcal{A}$ we have the initial rate of ILAD(iv). Step 2 relies on Lasso. Condition HL is implied by Condition I and Lemma 2 applied twice with $\zeta_i = v_i$ and $\zeta_i = d_i$ under the condition that $K_x^4\log p \leqslant \delta_n n$. By Lemma 6 we have $\|x'_i(\widehat{\theta} - \theta_0)\|_{2,n} \lesssim_P \sqrt{s\log(n \vee p)/n}$. Moreover, by Lemma 7 we have $\|\widehat{\theta}\|_0 \lesssim s$ with probability $1 - o(1)$.

The rates established above for $\widehat{\theta}$ and $\widehat{\beta}$ imply (A.25) in ILAD(iii) since by Condition I(ii) $\mathrm{E}[|v_i|] \leqslant (\mathrm{E}[v_i^2])^{1/2} = O(1)$ and $\max_{1\leqslant i\leqslant n} |x_i'(\widehat{\theta}-\theta_0)| \lesssim_P K_x\sqrt{s^2\log(p\vee n)/n} = o(1)$.

To verify Condition ILAD(iii) (A.26), arguing as in the proof of Theorem 1, we can deduce that

$$\sup_{\alpha\in\mathcal{A}} |\mathbb{G}_n(\psi_{\alpha,\widehat{\beta},\widehat{\theta}} - \psi_{\alpha,\beta_0,\theta_0})| = o_P(1).$$

This completes the proof.

□

Appendix G. Proof of Auxiliary Technical Results

Proof of Lemma 2. We shall use Lemma 8 ahead. Let $Z_i = (x_i, \zeta_i)$ and define $\mathcal{F} = \{f_j(x_i,\zeta_i) = x_{ij}^2\zeta_i^2 : j = 1,\ldots,p\}$. Since $\mathrm{P}(|X| > t) \leqslant \mathrm{E}[|X|^k]/t^k$, for $k = 2$ we have that $\mathrm{median}(|X|) \leqslant \sqrt{2\mathrm{E}[|X|^2]}$ and for $k = q/4$ we have $(1-\tau)$-quantile of $|X|$ is bounded by $(\mathrm{E}[|X|^{q/4}]/\tau)^{4/q}$. Then we have

$$\max_{f\in\mathcal{F}} \mathrm{median}\left(|\mathbb{G}_n(f(x_i,\zeta_i))|\right) \leqslant \sqrt{2\bar{\mathrm{E}}[x_{ij}^4\zeta_i^4]} \leqslant K_x^2\sqrt{2\bar{\mathrm{E}}[\zeta_i^4]}$$

and

$$(1-\tau)\text{-quantile of } \max_{j\leqslant p}\sqrt{\mathbb{E}_n[x_{ij}^4\zeta_i^4]} \leqslant (1-\tau)\text{-quantile of } K_x^2\sqrt{\mathbb{E}_n[\zeta_i^4]} \leqslant K_x^2(\bar{\mathrm{E}}[|\zeta_i|^q]/\tau)^{4/q}.$$

The conclusion follows from Lemma 8. □

Proof of Lemma 3. By the triangle inequality we have

$$\|x_i'(\widehat{\beta}^{(2m)}-\beta_0)\|_{2,n} \leqslant \|x_i'(\widehat{\beta}-\beta_0)\|_{2,n} + \|x_i'(\widehat{\beta}^{(2m)}-\widehat{\beta})\|_{2,n}.$$

Now let T^1 denote the m largest components of $\widehat{\beta}$ and T^k corresponds to the m largest components of $\widehat{\beta}$ outside $\cup_{d=1}^{k-1}T^d$. It follows that $\widehat{\beta}^{(2m)} = \widehat{\beta}_{T^1\cup T^2}$.

Next note that for $k \geqslant 3$ we have $\|\widehat{\beta}_{T^{k+1}}\| \leqslant \|\widehat{\beta}_{T^k}\|_1/\sqrt{m}$. Indeed, consider the problem $\max\{\|v\|/\|u\|_1 : v,u\in\mathbb{R}^m, \max_i|v_i| \leqslant \min_i|u_i|\}$. Given a v and u we can always increase the objective function by using $\tilde{v} = \max_i|v_i|(1,\ldots,1)'$ and $\tilde{u}' = \min_i|u_i|(1,\ldots,1)'$ instead. Thus, the maximum is achieved at $v^* = u^* = (1,\ldots,1)'$, yielding $1/\sqrt{m}$.

Thus, by $\|\widehat{\beta}_{T^c}\|_1 \leqslant \mathbf{c}\|\delta_T\|_1$ and $|T| = s$ we have

$$\begin{aligned}
\|x_i'(\widehat{\beta}^{(2m)}-\widehat{\beta})\|_{2,n} &= \|x_i'\textstyle\sum_{k=3}^K\widehat{\beta}_{T^k}\|_{2,n}\\
&\leqslant \textstyle\sum_{k=3}^K\|x_i'\widehat{\beta}_{T^k}\| \leqslant \sqrt{\phi_{\max}(m)}\sum_{k=3}^K\|\widehat{\beta}_{T^k}\|\\
&\leqslant \sqrt{\phi_{\max}(m)}\textstyle\sum_{k=2}^{K-1}\frac{\|\widehat{\beta}_{T^k}\|_1}{\sqrt{m}} \leqslant \sqrt{\phi_{\max}(m)}\frac{\|\widehat{\beta}_{(T^1)^c}\|_1}{\sqrt{m}}\\
&\leqslant \sqrt{\phi_{\max}(m)}\frac{\|\widehat{\beta}_{T^c}\|_1}{\sqrt{m}} \leqslant \sqrt{\phi_{\max}(m)}\mathbf{c}\frac{\|\delta_T\|_1}{\sqrt{m}}.
\end{aligned}$$

□

Appendix H. Auxiliary Probabilistic Inequalities

Let $Z_1, \ldots, Z_n$ be independent random variables taking values in a measurable space $(S, \mathcal{S})$, and consider an empirical process $\mathbb{G}_n(f) = n^{-1/2} \sum_{i=1}^n \{f(Z_i) - \mathrm{E}[f(Z_i)]\}$ indexed by a pointwise measurable class of functions $\mathcal{F}$ on S (see [25], Chapter 2.3). Denote by $\mathbb{P}_n$ the (random) empirical probability measure that assigns probability n^{-1} to each Z_i. Let $N(\epsilon, \mathcal{F}, \|\cdot\|_{\mathbb{P}_n,2})$ denote the ϵ-covering number of $\mathcal{F}$ with respect to the $L^2(\mathbb{P}_n)$ seminorm $\|\cdot\|_{\mathbb{P}_n,2}$.

The following maximal inequality is derived in [6].

Lemma 8 (Maximal inequality for finite classes)**.** *Suppose that the class $\mathcal{F}$ is finite. Then for every $\tau \in (0, 1/2)$ and $\delta \in (0, 1)$, with probability at least $1 - 4\tau - 4\delta$,*

$$\max_{f \in \mathcal{F}} |\mathbb{G}_n(f)| \leqslant \left\{ 4\sqrt{2 \log(2|\mathcal{F}|/\delta)}\ Q(1-\tau) \right\} \vee 2 \max_{f \in \mathcal{F}} \mathrm{median}\left(|\mathbb{G}_n(f)|\right),$$

where $Q(u) :=$ u-quantile of $\max_{f \in \mathcal{F}} \sqrt{\mathbb{E}_n[f(Z_i)^2]}$.

The following maximal inequality is derived in [3].

Lemma 9 (Maximal inequality for infinite classes)**.** *Let $F = \sup_{f \in \mathcal{F}} |f|$, and suppose that there exist some constants $\omega_n > 1$, $\upsilon > 1$, $m > 0$, and $h_n \geqslant h_0$ such that*

$$N(\epsilon \|F\|_{\mathbb{P}_n,2}, \mathcal{F}, \|\cdot\|_{\mathbb{P}_n,2}) \leqslant (n \vee h_n)^m (\omega_n/\epsilon)^{\upsilon m}, \ 0 < \epsilon < 1.$$

Set $C := (1 + \sqrt{2\upsilon})/4$. Then for every $\delta \in (0, 1/6)$ and every constant $K \geqslant \sqrt{2/\delta}$, we have

$$\sup_{f \in \mathcal{F}} |\mathbb{G}_n(f)| \leqslant 4\sqrt{2} c K C \sqrt{m \log(n \vee h_n \vee \omega_n)} \max\left\{ \sup_{f \in \mathcal{F}} \sqrt{\bar{\mathrm{E}}[f(Z_i)^2]},\ \sup_{f \in \mathcal{F}} \sqrt{\mathbb{E}_n[f(Z_i)^2]} \right\},$$

with probability at least $1 - \delta$, provided that $n \vee h_0 \geqslant 3$; the constant $c < 30$ is universal.

References

[1] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28:253–263, 2008.

[2] A. Belloni, D. Chen, V. Chernozhukov, and C. Hansen. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2430, November 2012.

[3] A. Belloni and V. Chernozhukov. ℓ_1-penalized quantile regression for high dimensional sparse models. *Annals of Statistics*, 39(1):82–130, 2011.

[4] A. Belloni and V. Chernozhukov. Least squares after model selection in high-dimensional sparse models. *Bernoulli*, 19(2):521–547, 2013.

[5] A. Belloni, V. Chernozhukov, and I. Fernandez-Val. Conditional Quantile Processes based on Series or Many Regressors. *ArXiv e-prints*, May 2011.

[6] A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection amongst high-dimensional controls. *ArXiv*, 2011.

[7] A. Belloni, V. Chernozhukov, and C. Hansen. Inference for high-dimensional sparse econometric models. *Advances in Economics and Econometrics. 10th World Congress of Econometric Society, held in August 2010*, III:245–295, 2013.

[8] A. Belloni, V. Chernozhukov, and K. Kato. Robust inference in high-dimensional sparse quantile regression models. *Working Paper*, 2013.

[9] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

[10] Victor Chernozhukov and Christian Hansen. Instrumental variable quantile regression: A robust inference approach. *Journal of Econometrics*, 142:379–398, 2008.

[11] Xuming He and Qi-Man Shao. On parameters of increasing dimensions. *J. Multivariate Anal.*, 73(1):120–135, 2000.

[12] Bing-Yi Jing, Qi-Man Shao, and Qiying Wang. Self-normalized Cramer-type large deviations for independent random variables. *Ann. Probab.*, 31(4):2167–2215, 2003.

[13] K. Kato. Group lasso for high dimensional sparse quantile regression models. Preprint, ArXiv, 2011.

[14] Roger Koenker. *Quantile regression*, volume 38 of *Econometric Society Monographs*. Cambridge University Press, Cambridge, 2005.

[15] Michael R. Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Series in Statistics. Springer, Berlin, 2008.

[16] Sokbae Lee. Efficient semiparametric estimation of a partially linear quantile regression model. *Econometric theory*, 19:1–31, 2003.

[17] Hannes Leeb and Benedikt M. Pötscher. Model selection and inference: facts and fiction. *Economic Theory*, 21:21–59, 2005.

[18] Hannes Leeb and Benedikt M. Pötscher. Sparse estimators and the oracle property, or the return of Hodges' estimator. *J. Econometrics*, 142(1):201–211, 2008.

[19] J. Neyman. Optimal asymptotic tests of composite statistical hypotheses. In U. Grenander, editor, *Probability and Statistics, the Harold Cramer Volume*. New York: John Wiley and Sons, Inc., 1959.

[20] J. Neyman. $c(\alpha)$ tests and their use. *Sankhya*, 41:1–21, 1979.

[21] J. L. Powell. Censored regression quantiles. *Journal of Econometrics*, 32:143–155, 1986.

[22] M. Rudelson and S. Zhou. Reconstruction from anisotropic random measurements. *ArXiv:1106.1151*, 2011.

[23] R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, 58:267–288, 1996.

[24] S. A. van de Geer. High-dimensional generalized linear models and the lasso. *Annals of Statistics*, 36(2):614–645, 2008.

[25] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics, 1996.

[26] Lie Wang. L_1 penalized lad estimator for high dimensional ilnear regression. *ArXiv*, 2012.

[27] Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low-dimensional parameters with high-dimensional data. *ArXiv.org*, (arXiv:1110.2563v1), 2011.

[28] S. Zhou. Restricted eigenvalue conditions on subgaussian matrices. *ArXiv:0904.4723v2*, 2009.